University of Gondar
College of Natural and Computational Sciences
Department of Statistics

Modeling Time to Death and Identifying Its Determinants among HIV/AIDS Patients under ART Follow-Up in Gondar Teaching Hospital, Ethiopia

By
Yesuf Abdela

Advisor: Dr. Salie Ayalew (PhD)
Department of Statistics

A Thesis Submitted to the Department of Statistics, College of Natural and Computational Sciences, University of Gondar in Partial Fulfillment for the Requirements of Master of Science (M.Sc.) Degree in Bio-Statistics

August, 2015
Gondar, Ethiopia

University of Gondar, College of Natural and Computational Science, Department of Statistics

In the role of research advisor of the candidate, I, the undersigned, hereby certify that I have read and evaluated the thesis prepared by Yesuf Abdela Mustefa under my guidance, which was entitled as ***Modeling Time-to-Death and Identifying Its Determinants among HIV/AIDS Patients under ART Follow-Up in Gondar Teaching Hospital***. I approved that the thesis be submitted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

Dr. Salie Ayalew (PhD)		/ /
Advisor	Signature	Date

Being the members of the board of examiners of M.Sc. thesis open defense examination of Yesuf Abdela Mustefa, we certify that we have read and evaluated the thesis and examined the candidate. We, the undersigned, approved that the thesis be accepted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

		/ /
Name of Chairman	Signature	Date
		/ /
Name of Main Advisor	Signature	Date
		/ /
Name of Internal Examiner	Signature	Date
		/ /
Name of External Examiner	Signature	Date

Gondar, Ethiopia

DECLARATION

As researcher of the thesis, I, the undersigned, assert that the thesis is my original work, has not been presented for degrees in any other University and all sources of materials used for the thesis have been duly acknowledged.

Yesuf Abdela Mustefa		/ /
Name	Signature	Date

Table of Contents

List of Tables

List of Figures

Acronyms

AFT	Accelerated Failure Time
AIC	Akaike's Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
AZT	Zidovudine (Retrovir)
BIC	Bayesian Information Criteria
BMI	Body Mass Index
CI	Confidence Interval
FDA	Food and Drug Administration
GG	Generalized Gamma
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
HSCH	Healthcare School of Hawaii
IQR	Interquartile Range
KM	Kaplan-Meier
LOWESS	Locally Weighted Scatterplot Smoothing
LR	Likelihood Ratio
ML	Maximum Likelihood
OIs	Opportunistic Infections
PH	Proportional Hazards
PHR	Proportional Hazards Regression
PMTCT	Prevention of Mother-To-Child Transmission
TB	Tuberculosis
TR	Time Ratio
UNAIDS	United Nations Program of HIV/AIDS
WHO	World Health Organization
ZDV	Zidovudine

Acknowledgement

First and foremost, my deepest gratitude goes to my advisor, Dr. Salie Ayalew, for his excellent professional advice and supervision was not separated from me until the completion of this thesis.

I would like very much to thank Dr. Asrat Atsedeweyn as former Department Head, for providing me recommendation and authentication letters that helped me get supports from university of Gondar Hospital, ART data clerk staff. The current Department Head, Aragaw Eshetie (M.Sc.) to whom that I am enormously grateful, also made principal contribution in forwarding up to date information.

It is my desire to acknowledge the help of all friends of mine, who supported me in any aspects of the study during its course.

Last but not least, my heart felt gratitude goes to my family especially my mother, Fatuma Selman, as well as my fiancé, Beydu Awol, for their uninterrupted moral and financial support that helped and extraordinarily encouraged me the whole period in the accomplishment of my research paper.

Abstract

Although remarkable decrease in the number of HIV related morbidity and mortality was recorded since the start of ART, exploring the effects of different factors determining survival of patients plays principal role So as to enhance ART. In such cases, survival analysis is the most appropriate method and its parametric version yields more powerful estimates than Cox PH provided that, distributional assumptions satisfied. Hence, the performances of Cox PH model and different parametric Models in both metric (PH and AFT) were compared on 3042 eligible subjects followed for at least six years using secondary data obtained from University of Gondar Hospital, ART database. Model specific stepwise variable selection procedures were performed on 11 Covariates that appeared to be significant by Non-Parametric and independent analysis to fit models with the corresponding set of covariates resulting possibly minimum loss of information. Based on the statistical information criteria, the Generalized Gamma AFT Model provided best fit to the dataset as compared to Cox PH, Exponential, Weibull, Gompertz, Log-Logistic and Log-Normal Models. Accordingly, advanced WHO clinical stages (III and IV), lower CD4 percent (12-15%), TB co-infection, being bedridden or ambulatory functional status, being relatively old in age, having relatively lower weight, the presence of Opportunistic Infections and Risky Behaviors were strongly related to relatively minimum median survival time (accelerated death time) under the Generalized Gamma AFT Model. Gender, Household Size and Educational Level were not selected to fit final Generalized Gamma AFT Model, although they were significant by independent analysis unlike Occupational Status, Marital Status and Cotrimoxazol that were insignificant at all. Statistical software, SPSS and STATA were used for analysis.

Key Terms: ART, Model Specific Stepwise Procedures, Generalized Gamma AFT Model.

1. Introduction

1.1. Background

The human immunodeficiency virus (HIV) has created an enormous worldwide challenge since its recognition. According to [71], it has been possible to know that in 2013 there were an estimated 35 million people living with HIV in the entire globe [26]. It had also shortened the lives of many others for at least three decades without cure. This pandemic is spreading around the globe. It is the reason for many children being orphaned. Since the start of the epidemic, 39 million people have died of AIDS related illness [71].

Globally, an estimated 2.1 million people were newly infected with HIV in 2013. This estimate appeared to be an indication of continues fall in the number of new HIV infections in the world because it was the lowest number of annual new infections since the mid-to-late 1990s, when approximately 3.5 million people were acquiring HIV every year [68]. The drop in new HIV infections was most pronounced among children. From 2001 to 2013 the number of children newly infected with HIV dropped by 58% [71].

Although decrease in number of new infection with HIV is remarkable event, HIV/AIDS still has devastating effect on lives of mankind. Considering this large number of loss in life due to HIV/AIDS, a lot has been done to reduce HIV/AIDS related mortality and morbidity. Now days, this contributed effort brought ART program which has been hope for survival of HIV infected patients since its announcement. On the other hand, high costs associated with treatments were reasons for losing hope to wards the treatments for HIV/AIDS right after the announcement of ARV drug combinations. Despite, it was believed to be better way of prolonging the life span of HIV infected patients. Fortunately, such problems seem to be solved at this time as the pandemic has gained global concern [72].

According to [71], the highest proportion of people living with HIV is saturated in sub-Saharan Africa. The estimated total number of people living with HIV in sub-Saharan Africa by the year 2013 was 24.7 million. It is also reported that a minimum of 1.4 million people were newly infected with HIV in 2012 and an estimated 1.1 million sub-Saharan people experienced HIV/AIDS related death in 2013 [66]. However, new HIV infection was declined by 33% between 2005 and 2013 [15].

Ethiopia is one of sub- Saharan Africa countries in which an interval estimate of [690,000- 840,000] people were living with HIV since 2012 [42], [72]. Since the first evidence of HIV was detected in Ethiopia in 1984 [69], AIDS has claimed the lives of millions and left behind hundreds of thousands of orphans [21], [22]. It has also reported that there were 14,000 to 29,000 new HIV infections and an interval estimate of [40,000-56,000] AIDS related deaths [72].

In 2003, the government of Ethiopia introduced ART program to reduce HIV related morbidity and mortality, improve the quality of life of people living with HIV and mitigate some of the impacts of the epidemic [28],[29]. Again in 2005, Ethiopia launched free ART: over 71,000 were initiated on ART by the end of November 2006 and as of March 2010, 511 health facilities (142 hospitals and 369 health centers) provide ART service throughout the country to increase access by taking service closer to more people recording transport and related costs for patients and families, resulting in improving adherence and enrollment in care and treatment services easily in the course of the disease [44], [73].

In order to achieve the goals of ART, it is better to study different socio-economic, demographic and health factors that are expected to have significant effect on the survival of HIV infected patients. Besides investigating ways of prevention of the virus before a person is HIV positive, it is also of recent interest in medical researches related to HIV/AIDS to identify marker events such as time to death, and determining factors affecting the length of survival time among HIV infected patients under ART follow-up.

Due to the presence of various determinants of time to death or survival time, there is a great deviation in time to death of HIV infected patients under ART follow-up. According to different studies, the survival time of HIV patients under ART follow-up is known to be influenced by socio-economic, demographic, and health factors.

Although different statistical methodologies can be considered to identify such influential factors on survival of individuals particularly, HIV patients under ART follow-up, survival analysis is more preferable method of statistical analysis to determine factors influencing life time of individual patients since time to death can be categorized as ‘time to event’ data[45]. In this particular thesis, individual survival times before experiencing

death outcome are defined as the length of time from ART start date until the date of death (or censor) measured in months.

By definition, Survival analysis is a phrase used to describe the analysis of data in the form of times from a well-defined time origin until the occurrence of some particular event or end point [10]. Hence, survival analysis is also referred to as "time-to-event analysis", which is applied in a number of applied fields, such as medicine, public health, social science, and engineering [34].

On the basis of assumptions, Non- parametric, semi- parametric and parametric methods are classification of statistical methods used to analyze a survival data. Non-parametric methods work well for homogeneous samples; they do not determine whether or not certain variables are related to the survival times [45]. This leads to the application of regression methods for analyzing survival data. The standard multiple linear regression models are not well suited to survival data for several reasons. First, survival times are rarely normally distributed. Second, censored data result in missing values for the dependent variable [45].

In survival analysis studies, a descriptive summary of characterizing the different survival distributions that correspond to different subgroups within a heterogeneous population could consist of parametric or semi parametric methods. There are two major regression models used for survival data: proportional hazards model (Cox) as a semi parametric method and accelerated failure time model or linear model representation in log time as a parametric model [45].

Cox proportional hazard model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the nature or shape of the baseline hazard function. That is why Cox’s model is referred to as a semi-parametric model for the hazard function. This model keeps the baseline hazard as an arbitrary, unspecified, and nonnegative function of time. It is the most popular and commonly used model by researchers in medical sciences mainly because of its simplicity, and not being based on any assumptions about the survival distribution [16]. However, Cox PH model has the restriction that proportional hazards assumption holds with time-fixed covariates;

and it may not be appropriate in many situations and other modifications such as stratified Cox model or Cox model with time-dependent variables are required [40].

The Accelerated Failure Time (AFT) model is another alternative method for the analysis of survival data. Although the Cox regression model is the most favorable employed technique in survival analysis, parametric models do have a number of benefits [45].

The details on different survival data analysis methods are presented on the 3rd chapter of this thesis. However, the main purpose of this study is to support an argument that AFT models can be used as best alternative to Cox PH models to determine factors that significantly affect the survival time of HIV patients.

1.2. Statement of the Problem

Although a lot of enquiries have been conducted in different hospitals of Ethiopia in investigating factors affecting the survival time of HIV infected patients under ART follow-up, there are inconsistencies in results of different studies which leaded to contradicting conclusions about the effect of some factors on survival times of patients. Specifically, age, gender and educational level are some of the factors that the same conclusions couldn't be given by different researchers regarding their effect on survival time of HIV patients under ART follow-up. Hence, degree of uncertainty associated with the effect of such factors needed to be minimized.

Moreover, almost all of the studies were conducted on patients of age between 15 and 75 years for which the maximum follow-up time was not more than 5 years. So that, conclusions can't be made for patients of age other than between 15 and 75. In addition to this, results of such studies may be affected by high proportion of censored observations due to relatively shorter follow-up time since death from HIV/AIDS take a decade on average unless patients die from other cases. Basically, it is now becoming many years since ART has been launched in Ethiopia. As a result, the time at which the studies have been conducted also has effect on the length of follow-up time because recently conducted studies could have relatively long follow-up time to consider than those had been conducted before. Number of observed survival times were considered in sample size determination rather than number of death events in some of the studies. Therefore, it was not only the length of follow-up time, but also the mechanism by which

samples have been taken was one of the main causes to have high proportion of censored survival times.

Essentially, appropriate method of statistical data analysis should be used to get accurate results. Some of the studies conducted so far used logistic framework. However, for time -to -event data, survival analysis method is more powerful than the logistic framework as it takes censuring into consideration while the probit analysis (logistic framework) restricted attention to the events that occur within the shortest observed follow-up time that leads to a huge waste of information.

Most importantly, the absence of studies conducted to compare parametric (AFT) with Cox PH methods of survival analysis on the basis of determining most significant factors affecting the length of life time of patients under ART follow-up especially in Ethiopia motivated this study since parametric methods of statistical analysis is more powerful than non-parametric and semi-parametric method of analysis provided that distributional assumptions are satisfied. Moreover, comparisons made before on datasets other than ART were performed on the basis of the same sets of covariates regardless of model specific stepwise procedures.

In this context, the following questions are addressed:-

- Does Cox PH model provides better fit to time-to-death HIV data than parametric survival (AFT) model in this particular study?
- Which socio-economic, demographic and health factors have statistically significant effect on survival time of HIV infected patients?
- Which factors/categories are associated with shorter survival time of HIV patient?
- Which factors/categories are associated with prolonged survival time of patients?

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study is to identify statistically significant determinants of time to death event of HIV infected patients under ART follow-up in Gondar hospital, Ethiopia.

1.3.2. Specific Objectives

- To identify factors that affects the survival time of HIV/AIDS patients under ART follow-up.
- To assess performance of Cox PH model relative to parametric (AFT) models based on a given particular dataset.
- To formulate a model that results in statistically plausible and interpretable estimates of the effect of important determinants of survival time for ART data.

1.4. Significance of the Study

- The outcome of this study may help concerned health policy makers to plan and design effective strategies and policies which can undertake children in to consideration to minimize HIV related mortality.
- It minimizes degree of uncertainty on the effect of some factors on survival time of HIV patients.
- It helps to identify the most important determinants of death among HIV infected patients so that appropriate possible interventions could be made by clinicians on an intent to control factors related to accelerated death time HIV/AIDS patients.
- It helps to examine the nature of correlation/association between survival time and different socio-economic, demographic and health factors by using both semi-parametric and parametric methods of survival analysis.
- It gives basic clue about fitting the different alternatives of modeling time to event data in survival analysis.
- It helps to investigate the impact of TB on HIV patients.
- It paves way for researchers to conduct further enquiry as it can contribute for growing literatures on statistical analysis particularly, on modeling time to event data; survival analysis.

1.5. Demarcation of the Study

1.5.1. Scope of the Study

The major concern of this study is determining the most important statistically significant socio-economic, demographic and health factors that can affect the life span of HIV/AIDS patients. It also categorize factors/factor levels to differentiate among which

categories of factors are associated with accelerated death time and which of them belong to prolonging life time of HIV infected patients.

Besides identifying those factors, both parametric and semi-parametric survival analysis methods are used in modeling death time of HIV/ AIDS patients on intent to support the argument that AFT models can be used as alternatives to Cox PH in modeling time to event data. In this regard, modeling time to event is modeling time to death of HIV infected patients under ART follow-up in Gondar hospital, Ethiopia. More over this, the non-parametric methods of survival analysis are used to describe the median survival time of patients. Cox PH and the best fitted AFT model(s) are also checked for assumptions.

All age groups are considered for this study including children. After determining eligible subjects for retrospective cohort of study, subjects were followed for at least six years right after the initiation of ART to experience death from any case or being censored.

1.5.2. Limitations

- The study presumed that all deaths were caused by HIV/AIDS.
- Parts of information on some individuals was missed due to the presence of drop out, transfer out, withdrawal or lost to follow-up.
- All attributes required for this study were not recorded either because of transfer in from other ART clinic or poor data base management system.
- Poor data searching system on different patient charts to identify event time.
- The study was based on baseline values of the variables of interest.

2. Review of Related Literatures

2.1. Theoretical Literature Review

2.1.1. Overview of HIV and AIDS

Although we often hear the terms “HIV” and “AIDS” used interchangeably, HIV and AIDS are not the same things. The abbreviation HIV stands for Human Immunodeficiency Virus. HIV is the virus that causes Acquired Immune Deficiency Syndrome, otherwise known as AIDS. AIDS develops in the late stages of HIV infection [38].

A person who has been infected with HIV is referred to as being HIV positive (HIV+). An HIV+ person may be asymptomatic, meaning that he may not have any symptoms of being infected. Although there are many negative stereotypes of HIV-infected people, it is impossible to look at someone and tell if he is HIV+. Many people who are infected with HIV may look and feel healthy [59]. Just because a person has been infected with HIV does not mean that he has or will be certain to develop AIDS. However, left untreated, most people with HIV infection eventually do develop AIDS. Without treatment, the time frame between a person becoming infected with HIV and subsequently developing AIDS is generally eight to ten years. However, there are cases of HIV+ people remaining asymptomatic for over two decades [38].

As noted, AIDS is an acronym for acquired immune deficiency syndrome. This is a condition that develops from HIV infection. The full name is an accurate description of what occurs to a body infected with HIV; namely, it results in a deficient immune system [38].

As most people know, our immune system is the body’s collective attempt to fight off whatever may be compromising healthy bodily functioning (bacteria, viruses, etc.). It is activated when we become ill or have the potential to become ill. For healthy people, it permits timely recovery and healing from illness. When compromised, it results in a reduced ability to fight off infection and increases the likelihood of becoming ill with other diseases. These infections are called “opportunistic” because they take the opportunity to attack when immune systems are immunosuppressed, meaning the immune system is functioning too poorly to fight off the infection. Although we

commonly hear and use the terminology of someone “dying of AIDS,” people actually do not die of AIDS. What they do die of are these opportunistic infections that overwhelm their bodies [38].

2.1.2. Origin of the Virus

The earliest known case of HIV infection was discovered in a blood sample drawn in 1959 from a man in Kinshasa, Democratic Republic of Congo (formerly Zaire). Although many theories about the origin of AIDS ranging from government conspiracies to aliens have been suggested, many researchers now generally agree that AIDS began somewhere in the Central African region. They believe that it likely developed from a simian virus that infected chimpanzees and somehow managed to cross over into humans, mutating into HIV perhaps in the 1930s. That could have happened when people ate the meat of, or were bitten by, the infected animals. Due to the geographic isolation of the area, the virus likely traveled out of the region slowly, eventually establishing itself in human hosts who spread the virus unaware of its existence [38].

2.1.3. Transmission of HIV

Like all other epidemic and pandemic diseases throughout history, when AIDS was identified, there was a great amount of fear about how it could be transmitted. Particularly early in the AIDS crisis, before any AIDS medications were available and at which time AIDS was considered always fatal, fears were especially profound. Social stigmas such as the prevalence of HIV/AIDS among marginalized groups such as gay men and intravenous drug users compounded the fears and resulted in discrimination. In some cases, people were so afraid that their behavior toward people who were infected with or even suspected to have HIV even became violent [38].

Researchers have now concluded that HIV is not an airborne virus like the influenza virus that means, someone cannot get infected with HIV, for example, by conversing with or sitting near someone in an airplane or theater who is HIV+. Fortunately, HIV is not a very robust virus when outside of the body. This means that it deteriorates fairly rapidly when not in an ideal environment like the body provides. It is neither able to replicate nor reproduce outside of the body, and any fluid that is infected with HIV that dries due to

exposure to the outside environment effectively renders HIV dead. This is why the likelihood of contracting HIV from casual environmental contact is remote [38].

HIV is not transmitted by insects such as fleas (as was the case in the plagues that swept through the world during the medieval period), and it is not transmitted by mosquitoes (as is malaria or the West Nile Virus). Additionally, HIV is not transmitted by touching, hugging, or shaking hands with an infected person [38].

HIV is transmitted through four body fluids: blood; semen; vaginal secretions; and breast milk. As such, prevention efforts have been, and continue to be, focused on limiting or eliminating the possibility of individuals transferring these fluids between one other. This is why there have been such well known public campaigns for safe (or safer) sexual practices, syringe exchange programs, and HIV testing for pregnant women; these campaigns have targeted the primary means of HIV transmission as a collective effort to reduce the likelihood of transmission [38].

Other unusual cases of HIV transmission have been documented, but upon investigation, all have involved some exchange of one of these infected body fluids [64]. For example, in a rare case of transmission by deep kissing, both people had bleeding gum disease [8]. In other rare cases in which transmission occurred from adult to child from HIV+ adults pre-chewing food for infants, the adults also had bleeding gum disease and fed teething children [27].

2.1.4. How Does HIV Work?

Recall that HIV is the acronym for the human immunodeficiency virus. Notice that the final word in this sentence is virus; this is precisely what HIV is, a virus. Viruses are microscopic biological agents that are technically not considered to be living, as they do not meet the scientific requirements for what constitutes life (e.g., able to grow and reproduce, adapt to environmental conditions, etc.) They are only able to replicate and reproduce themselves through the use of other cells. Specifically, they attach (infect) themselves to a host cell and deliver their genetic material to the cell. They then hijack the cell's mechanisms for reproduction and use the cell to replicate many versions of themselves. Eventually, the cell becomes full of the replicated viruses and its structure begins to fail. The volume of replicated viruses in the cell causes the cell to burst,

destroying the cell and releasing the replicated viruses to infect other cells. In this manner, the viruses spread to other cells replicating themselves, and in the process they destroy all of the cells used as replication centers [38].

HIV attaches to two types of white blood cells: T cells and CD4 cells. These cells are components of the human immune system, and their cellular health is vital to the health of the immune system. When HIV attaches to T-cells and uses them as hosts to create more copies of it, it destroys them in the process. Destruction of T-cells results in a critically impaired immune system; this, then is what leads to the condition known as AIDS [38].

The Human Immunodeficiency Virus (HIV), a retrovirus, was identified in 1983 as the etiologic agent for the Acquired Immunodeficiency Syndrome (AIDS). AIDS is characterized by changes in the population of T-cell lymphocytes that play a key role in the immune defense system. In the infected individual, the virus causes a depletion of subpopulation of T-cells, called T-helper cells, which leaves these patients susceptible to opportunistic infections as well as certain malignancies [8], [47].

So, HIV is simply the virus with which one gets infected. Preventing AIDS requires preventing HIV infection. This is why it is important to understand how HIV is contracted, how it spreads, who is most at risk, what kinds of prevention practices can be put in place, etc. If HIV is prevented, the development of AIDS is prevented [38].

As noted above, unlike infection from other viruses that result in changes in homeostasis in the body, infection with HIV does not immediately result in symptoms like chills, fever, aches, and pains, etc. Persons infected with HIV can live symptom free for many years; in fact, it is estimated that 25% of people infected with HIV are unaware of being infected. The only way to know if someone is infected is to get tested. This is why there has been an effort for the past 25 years for people to get tested for HIV [38].

2.1.5. Developing Drugs to Treat HIV

Before HIV was isolated as the virus that causes AIDS, doctors were at a loss for effective ways to help their patients. They treated the opportunistic infections, but could

do nothing about addressing the cause of AIDS, so AIDS was widely considered a death sentence. Desperate patients grasped at desperate measures [38].

Earlier researchers looking for a drug to combat HIV focused on experiments using known antiviral medications. These drugs generally did little to benefit patients; some had uncomfortable of even dangerous side effects, but they were the only hope many people with HIV and their loved ones had. When Ribavirin, an established drug used for viral respiratory disease, showed some potential in slowing the onset of AIDS, AIDS patients illegally imported it from Mexico before being able to obtain it use against AIDS in the United States [20].

The first anti-HIV drug approved by the FDA as effective in fighting AIDS was azidothymidine, commonly known as AZT, and sold as Retrovir. It was approved for use by the FDA on March 19, 1987. Although the breakthrough in treatment was exciting, AZT therapy was cumbersome. It required some 12 pills per day, taken two at a time every four hours around the clock [63]. It was also extremely expensive: $3.00 per pill, or $8,000–$10,000 for a typical year's supply of medication. Public pressure eventually led to decreased costs. Even so, the drug was extremely profitable for the manufacturer, earning more than $300 million annually by 1994 [20], [60]. Even the WHO in a 1994 meeting declined to endorse AZT for use around the globe for pregnant women due to costs and access issues, recommending a solution of "simpler and less costly" therapies be developed [60]. AZT was also proved to have uncomfortable and even dangerous side effects as well (e.g., nausea, vomiting, headaches, fatigue, anemia, muscle pain and weakness, and neutropenia a low white blood cell count that increases susceptibility to infection) to the extent that some patients stopped taking the drug.

Researchers raced to develop new drugs to fight HIV, yet the virus was proved to be a formidable adversary. Researchers discovered that problems arose with so-called mono therapy treating HIV with only one drug. When attacked with only one drug at a time, HIV mutated and became resistant to that drug. The resistant strains of HIV could then be passed on to others [38].

As this problem became increasingly apparent to researchers and newer drugs became available, standard therapy for HIV starting in the mid-to-late 1990s became combination

therapies that used various drugs together to combat HIV. Highly Active Anti-Retroviral Therapy, known by the acronym HAART, uses combinations of three or more drugs to reduce the chance of HIV drug resistance. Doctor David Ho of New York City's Aaron Diamond AIDS Research Center was recognized as Time magazine's Person of the Year for 1996 for his pioneering work on combination therapies [9], [20]. However, problems with side effects, patient adherence to difficult medication schedules (some requiring 30 pills to be taken throughout the day with varying dietary requirements), and high drug costs persisted [20]. By 2008, over 25 different combinations of medications had been developed as "second line" therapies and even as "salvage" therapies to fight these drug resistant strains. One pill a day therapies multidrug combination products have also been developed to try to overcome some of the ongoing problems of resistance, side-effects, and cumbersome medication schedules [18], [24], and [64].

Pharmaceutical development is a lengthy and expensive process involving a number of stages. In the early stages, the drug is developed and tested in a laboratory and on animals. If the drug is determined to be safe and produces promising results, it goes through a process requiring three more phases of testing on human volunteers. The next phases are called clinical trials. These phases take several years (for example, more than two years for the first phase and up to four for the third phase) and include larger numbers of people for testing at each stage of the process, ranging from 10 to 100 volunteers in Phase I, to several hundred in Phase II, to several thousand in Phase III. Researchers progressively build data on the safety, effectiveness, dosage, and any side effects. If the drug proves satisfactory, it may be licensed and become available for widespread use [23], [37], [62], and [46]. Due to activism by HIV/AIDS advocacy groups, this process has actually been speeded up for HIV drugs.

2.2. Empirical Literature Review

2.2.1. Socio-Economic, Demographic and Health Factors Affecting Survival of HIV/AIDS Patients under ART Follow-Up

In this section of literature review, the main focus is to get information about socio-economic, demographic and health factors that are significant determinants of death time among HIV/AIDS patients under ART follow-up from previously conducted related studies. Moreover, this literature review also shows that some factors on which

contradicting conclusions have been made concerning the significance of their effect on survival.

According to a retrospective cohort study conducted in the Far-North province of Cameroon to analyze the outcomes of ART in routine conditions in a rural hospital on 1187 patients of age greater than 15 years who started ART between July 2001 and December 2006, CD4 count, hemoglobin, BMI, sex and clinical stage at enrolment were found to be independent predictors of mortality. In this study, the survival time was estimated by Kaplan–Meier analysis and Cox proportional hazard models were fitted to explain survival. Results Upon enrolment, 90.4% patients were in WHO stage III or IV and 56.1% had a BMI<18.5. Median CD4 count was 105 cells⁄mm3 (IQR 40–173). At the end of the study period, 338⁄1187 had died and 59⁄1187 were lost to follow-up. The survival probability was 77% at 1 year [95% CI: 75–80] and 47% at 5 years [95% CI: 40–55]. The median survival time was 58 months [31].

Similarly, another study was conducted based on data collected during the follow-up time from 2005 to 2008 at Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia. Out of a population of HIV patients who were taking antiretroviral therapy in the hospital in that period, data on 1,000 patients were used for this study. The study subjects were people in the age range from 15 to 75 years. The Kaplan-Meier Method was employed to estimate mortality; the Cox Proportional Hazards Regression Method was used to identify determinants of mortality. After initiation of the antiretroviral treatment, HIV-positive patients lived for an average of 5.65 years (CI:3.69-7.61 years); the median survival age was found to be 3.98 years (CI: 2.98-4.97 years). The number of medications, baseline functional status, CD4 count, antiretroviral treatment, age, gender and weight impact the survival experience of the patients [67].

On the other hand, the result of a study conducted in Adama Hospital, Ethiopia showed that age and access to running water appeared to have non-significant effect on survival/death status. The study evaluates factors affecting the chance of survival/death status among HIV positive people under ART follow-up. The socio economic factors such as previous HIV counseling, residence, employment status, number of rooms, availability of running electricity; demographic and health factors like gender, marital

status, educational level, TB status, weight, CD4 count and clinical stage; risk behavior factors:-condom use, tobacco alcohols and drug use were factors that had statistical significant effect on survival/death status of patients by multiple logistic regression model [50].

In contrast, according to a study conducted in south Wello, Ethiopia, It has determined that, the level of marital status, residence, TB status, clinical stage, weight, age, CD4 level, and lymphocyte count are that manifest differences in survival. But gender and education level show no significant effect on survival. In this study, a sample of 654 out of 7163 patients of age between 15 and 75 was selected that were followed from 1 January, 2008 to 31 December, 2011to identify factors related to the survival of HIV/AIDS patients under ART follow-up. The result of the study shows that, HIV-positive patients lived for an average of 41.81 months (CI: 40.61-43.00 months. The study employed The Kaplan-Meier estimator (product-limit-estimator) of the survival function to compare the survival functions of two or more groups. The log-rank test was utilized to test whether observed differences in survival experience between/among the groups was significant or not. The semi-parametric regression model known as the proportional hazards regression (PHR) model was also used for the analysis [33].

Furthermore, Educational level of individual patients appeared to have significant effect on survival in a study conducted in armed forces general teaching hospital Addis Ababa, Ethiopia. The study revealed that employment status, number of rooms, household size, functional status, CD4 cell count, WHO clinical stage, OIs, TB, ART, age and weight of the patients were also candidate predictors for further analysis. In this study patients were followed up for a median of 38.5 months and the overall mean estimated survival time of patients under the study was 72 (95% CI: 70- 74) months. Kaplan-Meier and Nelson Aalen estimation techniques were used to get a closer look at estimate of the survival time. In order to study the relationship between survival time and covariates, a regression modeling approach to survival analysis using the Cox proportional hazards model was employed [39].

In summary, the studies have been conducted to determine different factors that affect survival of HIV/AIDS patients under ART follow-up. However, different conclusions

have been made on the effect of some factors specifically, age, gender and educational level. The follow-up time for most of the studies employed survival analysis was not more than 4 years. And hence, analyses were affected by relatively high proportion of censored observations. Finally, most of the studies were conducted on specific age group of patients (>15 or between15 and 75). As a result, some of the researchers listed it as their limitation since results are not applicable for patient population of age excluded from the studies.

2.2.2. Comparing Cox PH versus AFT Survival Analysis Models

A review of literature on survival analysis used in different journals reveals that the Cox PH model is the most widely used way of analyzing survival data in clinical research. Researchers in medical sciences often tend to prefer semi-parametric instead of parametric models because of fewer assumptions. However, in recent time, AFT models as parametric models have attracted considerable attention, because not only they do not need PH assumption but also thanks to availability of standard statistical software ML parameter estimation and testing can be done readily [2],[45].

The main drawback of parametric models is the need to specify the distribution that most appropriately mirrors that of the actual survival times [14]. This is an important requirement that needs to be verified and an appropriate distribution may be difficult to identify. When a suitable distribution can be found, the parametric model is more informative than the Cox model. It is straightforward to derive the hazard function and to obtain predicted survival times when using a parametric model, which is not the case in the Cox framework. Additionally, the appropriate use of these models offers the advantage of being slightly more efficient; they yield more precise estimates (i.e. smaller standard errors) and that in the parametric models we often use maximum likelihood procedures to estimate the unknown parameters in which this technique and its interpretation are familiar for researchers [7].

The parametric approach offers more in the way of predictions, and the AFT formulation allows the derivation of a time ratio, which is arguably more interpretable than a ratio of two hazards in Cox PH models. However, AFT models are relatively unfamiliar and seen rarely in medical research [45].

To the best of my knowledge, many studies have not been conducted in comparing the performance the Cox PH and parametric AFT models using ART HIV/AIDS data to determine factors affecting the survival time of HIV/AIDS patients. However, the performance of Cox PH and AFT models has been compared in the analysis of HIV/TB Co-infection Survival data in Ghana. In this study, there were 76 patients on treatment of HIV/TB Co-infection from the year 2008 to 2013.the Cox model and (Exponential, Weibull, Lognormal, Log-logistics and Gamma) as AFT models have been compared using HIV/TB Co-infection data. The result showed that, the Cox model was fitted and diagnosed with the proportionality assumption satisfied. The martingale residual indicated that the model was linear. Comparing the Cox model with the AFT models based on the AIC and BIC showed that the Gamma model had the lowest value. It was also observed that weight, CD4 cell count and the Religion were significant determinants of the patient's survival at 5% significance level. The result revealed that the gamma model provided a better fit to the studied data than the Cox proportional hazards model [17].

Performance comparison between Cox PH and parametric survival models have also been made on other data sets other than ART such as survival of patients with different disease for example cancer. Accordingly, the analysis of Survival of Patients with Gastric Carcinoma of data from a historical cohort study of southern Iran with a diagnosis of stomach cancer has been made using Cox PH. and parametric Lognormal, Exponential, Gompertz, Weibull, Log logistic and Gamma regression models in which all parametric survival models were performed better than the Cox model. In this study the proportional assumption is checked and found to be hold, but the model diagnostic for the parametric case has not been made yet. The comparison of Parametric and semi parametric models was made based on AIC [54].

According to a study on Infant Feeding Survival and Markov Transition Probabilities among Children under Age 6 Months in Uganda, The multivariate analysis of the Cox, Weibull, and exponential models yielded similar results. Effect estimates were slightly lower in parametric models than in Cox models but were not significantly different for any of the parameters. Both the Weibull model and the exponential model had a lower Akaike Information Criterion than the Cox model and hence had better fits. A graphical

check of the proportional hazards assumption was performed, and the proportional hazards assumption did not appear to be violated (not shown). Graphical checks were also performed for all models using Cox-Snell residuals to determine model fit. These showed that both parametric models were better-fitting than the Cox model [41].

Another study conducted on retrospective cohort study of 746 patients with gastric cancer in Iran with the aim of comparing the Cox regression model against parametric. According to this study a multivariate analysis Cox and Exponential are similar. Although it seems that there may not be a single model that is substantially better than others, in univariate analysis the data strongly supported the log normal regression among parametric models and it can be lead to more precise results as an alternative to Cox [53].

Similarly, a study conducted on prognostic factors of survival time after hematopoietic stem cell transplant in acute lymphoblastic leukemia on 206 patients were enrolled after HSCH in Shariati Hospital between 1993 and 2007 so that the performance among AFT and Cox's models was assessed using explained variation and goodness of fit methods. Accordingly, predictive power of Weibull AFT models was superior to Cox PH model. Cox-Snell residual shows Weibull AFT fitted to data better than other distributions in multivariate analysis [57].

In a similar fashion, a study on the survival of 1236 tuberculosis patients admitted in randomized controlled clinical trial in India. The result for this study showed that AFT model gave smaller deviance showing that AFT models have seem to be more appropriate modeling than the Cox PH model [52].

Furthermore, AFT model was used to analyze data from 16 survivorship experiments in aging research experiments that evaluated the effects of one or more genetic manipulations on mouse lifespan. According to this study, AFT model deceleration factors also provided a more intuitive measure of treatment effect than the hazard ratio, and were robust to departures from modeling assumptions [65].

In contrast, in analysis of survival in acute severe illness, AFT models identified the same predictors as the Cox model, but did not demonstrate convincingly superior overall fit than the Cox PH model does and the analysis of survivals of breast cancer relapse

time with different treatments consistent results were obtained from accelerated failure time model and Cox proportional hazard model. But Cox PH is also chosen over accelerated failure time model to calculate the appropriate survival curves of relapse time for patients in different treatment groups. i.e., With respect to predicting survival curve, Cox-PH model gives better fit than AFT models [11].

3. Data and Research Methodology

3.1. Study Area and Data Source

In this retrospective cohort study, the information was collected from ART database in University of Gondar Referral Hospital. The University of Gondar Referral Hospital is found in Gondar city, which is located 727 km northwest of Addis Ababa. The ART service for University of Gondar Referral Hospital was initiated in 2003 and has had Adult, Pediatric, and PMTCT clinics [6]. The collected information was secondary data type from which necessary attributes have been obtained to identify the most important determinants of survival among HIV/AIDS patients under ART follow-up in Gondar hospital.

Intending to achieve the objective of this study leaded to the sampling frame HIV positive patients who were found in Gondar hospital under the follow-up of ART. The study focused on patients of all age who started ART between 2003 and 2009 that were followed until April, 2015 considering all possibly available attributes in the ART database. The obtained data was managed and analyzed by using statistical packages such as SPSS and STATA after the required information has been reviewed form ART database in Gondar hospital.

3.1.1. Study Variables

The Response Variable: In this particular study, the response (dependent) variable was survival time of HIV/AIDS patients under ART follow-up. The survival time is defined to be the length of time measured in month from ART initiation until time of death (censor).

Independent Variables: Predictors (covariates) taken at baseline value which were expected to have effect on the survival time of HIV/AIDS patients are listed as follows:-

- Age in years
- Sex (Male, Female)
- Marital status (Never Married, Married, Widowed or Divorced)
- Baseline Weight (kg)
- Level of education (Educated, Not Educated)

- Risk Behavior (Smoking, Alcohol, etc.) (No, Yes)
- Baseline CD4 cell Percent (12-15%, 16-28%, Above28%)
- WHO clinical stage (Stage I, Stage II, Stage III, Stage IV)
- Functional Status (Working, Ambulatory, Bedridden)
- Occupational status (Employed, Unemployed)
- Opportunistic infections (No, Yes)
- TB Status (Negative, Positive)
- Household size (Number of individuals per house)

3.2. Methods of Data Analysis

The methodology in this thesis has been built in such a way that non-parametric, semi-parametric and parametric methods of survival analysis could be applied to ART data on intent to compare the performance of Cox PH and AFT models in determining significant factors that affect survival time of HIV infected patients.

3.2.1. Introduction to Survival Analysis

Survival analysis is the phrase used to describe the analysis of data in the form of times from a well-defined time origin until the occurrence of some particular event or end point. Typically, Survival Analysis focuses on time to event data. In the most general sense, it consists of techniques for positive valued random variables [19]. In medical research, the origins will often correspond to the recruitment of an individual into an experimental study, such as clinical trial to compare two or more treatments. This in turn may coincide with the diagnosis of a particular condition, the commencement of a treatment regimen, or the occurrence of some adverse event. If the end point is the death of a patient, the resulting data are literally survival times. However, data of similar form can be obtained when the end point is not fatal, such as the relief of pain, or the reoccurrence of symptoms.in this case, the observations are often referred to as time to event data [10],[32].

The problem of analyzing time-to-event data arises in several applied fields such as medicine, biology, public health, epidemiology, engineering, economics, sociology, demography and etc. The terms lifetime analysis, duration analysis, event-history analysis, failure-time analysis, reliability analysis, and transition analysis refer essentially

to the same group of techniques although the emphases in certain modeling aspects could differ across disciplines [1], [51].

The main feature of survival data that renders standard methods inappropriate is that survival times are frequently censored [25]. The survival time of an individual is said to be censored when the end point of interest has not been observed for that individual. This may be because the data from a study are to be analyzed at a point in time when some individuals are still alive [55]. Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been lost to follow-up [10].

There are three types of censoring in survival analysis

A. Right Censoring: A patient who entered at a study time t_0 dies at time $t_0 + t$. However, t is unknown, either because the individual is still alive or because he/she has been lost to follow up. If the individual was last known to be alive at time $t_0 + c$, the time c is called a censored survival time. This censoring occurs after the individual has been entered in to a study. That is, to the right of last known survival time, and is therefore known as right censoring. The right censored survival time is then less than the actual, but unknown, survival time. This type of censoring is most commonly encountered in survival analysis and hence the term "censoring" will be used in this study to mean in all instances "right censoring".

B. Left Censoring: Another form of censoring is left censoring, which is encountered when the actual survival time of an individual is less than that observed. In other words, Survival time is said to be left censored if an individual develops an event of interest prior to the beginning of the study; left censoring occurs far less commonly than right censoring.

C. Interval Censoring: Here, individuals are known to have experienced an event with in an interval of time. That means, Survival time is said to be interval censored when it is only known that the event of interest occurs within an interval of time but the exact time of its occurrence is not known.

An important assumption for methods presented in this thesis for the analysis of censored survival data is that the individuals who are censored are at the same risk of subsequent

failure as those who are still alive and uncensored. i.e., a subject whose survival time is censored at time C must be representative of all other individuals who have survived to that time. If this is the case, the censoring process is called non-informative. Statistically, if the censoring process is independent of the survival time, then we will have non-informative censoring. Independence censoring is a special case of non-informative censoring. In this thesis, we assume that the censoring is non-informative right censoring.

3.2.2. Survivor Function and Hazard Function

In summarizing survival data there are two functions of central interest namely, survivor function and hazard function. In this section, definitions are given according to [10].

The actual survival time of an individual, t, can be regarded as the value of a variable, T, which can take any non-negative value. The different values that T can take have a probability distribution, and we call T the random variable associated with the survival time. Now suppose that the random variable T has a probability distribution with underlying probability density function $f(t)$. The distribution function of T is then given by

$$F(t) = P(T < t) = \int_0^t f(u)du,$$

And represent that the probability that the survival time is less than some value t. the survivor function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, and so

$$S(t) = P(T \geq t) = 1 - F(t). \qquad [3.1]$$

The survivor function can therefore be used to represent the probability that an individual survives from the time origin to some, time beyond t.

The hazard function is widely used to express the risk of hazard of death at some time t, and is obtained from the probability that an individual dies at time t, conditional on he/she having survived to that time [43]. For a formal definition of the hazard function, consider the probability that the random variable associated with an individual's survival time, T, lies between t and $t + \delta t$, conditional on T being greater than or equal to t, written $P(t \leq T < t + \delta t \mid T \geq t)$. This conditional probability is then expressed as a probability per

unit time by dividing by the time interval, δt, to give rate. The hazard function, $h(t)$, is then the limiting value of this quantity, as δt tends to zero, so that

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \le T < t + \delta t \mid T \ge t)}{\delta t} \right\} \qquad [3.2]$$

The function $h(t)$ is also referred to as the hazard rate, the instantaneous death rate, the intensity rate, or force of mortality. From equation [3.2], $h(t)\delta t$ is the approximate probability that an individual dies in interval $(t, t + \delta t)$, conditional on that person having survived to time t.

Again, from the definition of the hazard function in equation [3.2], we can obtain some useful relationships between the survivor and hazard functions. The conditional probability in the definition of the hazard function in equation [3.2] is

$$\frac{P(t \le T < t + \delta t)}{P(T \ge t)} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

Where $F(t)$ is the distribution function of T. Then, $h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}$

Now, $\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\}$ is the definition of the derivative of $F(t)$ with respect to t, which is $f(t)$ and so

$$h(t) = \frac{f(t)}{S(t)} \qquad [3.3]$$

It then follows that

$$h(t) = -\frac{d}{dt} \{\log S(t)\}, \qquad [3.4]$$

And so,

$$S(t) = \exp\{-H(t)\}, \qquad [3.5]$$

$$\text{Where } H(t) = \int_0^t h(u)du \qquad [3.6]$$

The function $H(t)$ defined in equation [3.6] features widely in survival analysis, and is called the integrated cumulative hazard. From equation [3.5], the cumulative hazard can be obtained from the survivor function, since

$$H(t) = -\log S(t). \qquad [3.7]$$

3.2.3. Non-Parametric Procedures

An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times for individuals in a particular group. Such summaries may be of interest in their own right, or as a precursor to a more detailed analysis of data. Survival data are conveniently summarized through estimates of the survivor function and hazard function. Methods for estimating these functions from a single sample of survival data are said to be non-parametric or distribution free, since they do not require specific assumptions to be made about the underlying distribution of the survival times [10].

3.2.3.1.Estimation of Survivor and Hazard Functions

In practice, when using actual data, we usually obtain estimated survivor function and curves that are step functions, rather than smooth curves. Among the other estimators of the survivor function the Kaplan-Meier estimator is the most common one. The Kaplan-Meier estimator of the survivorship function [Kaplan and Meier (1958)] also called product limit estimator, is the estimator used by most software packages. This estimator incorporates information from all of the observations available, both uncensored and censored, by considering survival to any point in time as a series of steps defined by the observed survival and censored times [30].

Assume we have a sample of n independent observations, their survival times denoted by $t_1, t_2, t_3, ...t_n$ and indicator of censoring by $\delta_1, \delta_2, \delta_3, ...\delta_n$ where $\delta_i = 1$, if an event/death occurs and $\delta_i = 0$ otherwise. Thus, the survival data are denoted by (t_i, δ_i) $i = 1,2,3,....n$. The first step to obtain the Kaplan-Meier estimator of the survival function is to order the survival times as $t_1 \le t_2 \le t_3 \le ..... \le t_n$. Assume that among the n observations $m \le n$ failures occurred at distinct m times. Then the rank- ordered failure times are $t_1 \le t_2 \le t_3 \le ..... \le t_m$.

Let n_i = the number at risk of dying or failure at t_i; d_i = the number of failures (deaths) at t_i. Then the Kaplan-Meier estimator of the survival function at time t is obtained from the equation,

$$\hat{S}_{KM}(t) = \prod_{t(_i)\le t}\left(\frac{n_i - d_i}{n_i}\right) = \prod_{t(_i)\le t}\hat{P}_i \qquad [3.8]$$

With the convention that $\hat{S}_{KM}(t) = 1$ if $t < t_{(1)}$ Thus,

$$\hat{H}_{KM}(t) = -\sum_{t_{(i)}\le t}\log\hat{P}_i \qquad [3.9]$$

The variance of the Kaplan-Meier estimators which is referred to as Greenwood's formula is given as:

$$\hat{Var}\left(\hat{S}_{KM}(t)\right) = \left[\hat{S}_{KM}(t)\right]^2\sum_{t_{(i)}\le t}\frac{d_i}{n_i(n_i - d_i)} \qquad [3.10]$$

And $\hat{Var}\left(\hat{h}(t)\right) = \left[\hat{h}(t)\right]^2\left(\frac{n_i - d_i}{n_i d_i}\right)$ [3.11]

Similarly, [1], [48], [49], and [3] have proposed an alternative estimator of $H(t)$that refers to as the Nelson-Aalen estimator. It is formulated as:

$$\tilde{H}(t) = \sum_{t_{(i)}\le t}\frac{d_i}{n_i} \text{ And it implies } \tilde{S}(t) = e^{-\tilde{H}(t)} = \prod_{t_{(i)}\le t}\exp\left(-\frac{d_i}{n_i}\right) \qquad [3.12]$$

It is merely in the case of small samples that the Nelson-Aalen estimate of the survivor function prevails over the KM estimate [30].

3.2.3.2.Estimating Median and Percentiles of Survival Times

Since the distribution of survival time tends to be positively skewed, the median is preferred for a summary measure of the location of the distribution. The median survival time is the time, beyond which 50% of the individuals in the population under study are expected to survive, and is given by the value $t(50)$ at which is such that $S\{t(50)\} = 0.5$ [10].

The estimated median survival time is given by $\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) < 0.5\}$. Where t_i is the observed survival time for the i^{th} individual, $i = 1,2,3,....., n$. In general, the estimate of the p^{th} percentile is $\hat{t}(p) = \min\left\{t_i \mid \hat{S}(t_i) < 1 - \frac{p}{100}\right\}$.

Approximate confidence intervals for the median and other percentiles of a distribution of survival times can be found once the variance of the estimated percentile has been obtained. An expression for the approximate variance of a percentile can be derived from a direct application of the general result for the approximate variance of function of random variable in the following equation. The variance of a function $g(x)$ of the random variable X is given by $Var\{g(x)\} \approx \left\{\frac{dg(x)}{dx}\right\}^2 Var(X)$. Using this result,

$$Var\left[\hat{S}\{t(p)\}\right] = \left(\frac{d\hat{S}\{t(p)\}}{dt(p)}\right)^2 Var\{t(p)\} = \left[-\hat{f}(t(p))\right]^2 Var\{t(p)\}$$

$$\Rightarrow Var\{t(p)\} = \left(\frac{1}{\hat{f}(t(p))}\right)^2 Var\left[\hat{S}\{t(p)\}\right]$$

Where, $t(p)$ is the p^{th} percentile of the distribution and $\hat{S}\{t(p)\}$ is the Kaplan-Meier estimate of the survivor function.

The standard error of $\hat{t}(p)$ is therefore given by $SE\left[\hat{t}(p)\right] = \frac{1}{\hat{f}(t(p))} SE\left[\hat{S}(\hat{t}(p))\right]$

The standard error of $\hat{S}(t(p))$ can be obtained using Greenwood's formula, given in equation [3.10], while an estimate of the probability density function at $\hat{t}(p)$ is

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{S}\{\hat{u}(p)\} - \hat{S}\{\hat{l}(p)\}}{\hat{l}(p) - \hat{u}(p)},$$

Where $\hat{u}(p) = \max\left\{t_{(j)} \mid \hat{S}(t_{(j)}) \geq 1 - \frac{p}{100} + \varepsilon\right\}$, and $\hat{l}(p) = \min\left\{t_{(j)} \mid \hat{S}(t_{(j)}) \leq 1 - \frac{p}{100} - \varepsilon\right\}$,

for $j = 1,2,3,..., r$, and small value of ε

In many cases, taking $\varepsilon = 0.05$ will be satisfactory, but a larger value of ε will be needed if $\hat{u}(p)$ and $\hat{l}(p)$ turn out to be equal. Therefore, for median survival time, $\hat{u}(50)$ is

the largest observed survival time from the K-M curve for which $\hat{S}(t) \geq 0.55$ and $\hat{l}(50)$ is the smallest observed survival time from the K-M curve for which $\hat{S}(t) \leq 0.45$. The 95% confidence interval for the p^{th} percentile $\hat{t}(p)$ has limits of $\hat{t}(p) \pm 1.96 * SE[\hat{t}(p)]$.

3.2.3.3.Non-Parametric Comparison of Survivorship Functions

In clinical research one is concerned not only with estimating the survival function but, more often, with the comparison of the life experience of two or more groups of subjects differing for a given characteristic or randomly allocated to different treatments. After providing a description of the overall survival experience in the study, we usually turn our attention to a comparison of the survivorship experience in key subjects in the data. The simplest way of comparing the survival times obtained from two or more groups is to plot the Kaplan-Meier curves for these groups on the same graph [56]. However, this graph does not allow us to say, with any confidence, whether or not there is a real difference between the groups. The observed difference may be a true difference, but equally, it could also be due to chance. Assessing whether or not there is a real difference between groups can only be done, with any degree of confidence, by utilizing statistical tests. Since survival data are typically right skewed, we would likely use rank-based non-parametric tests followed by estimates and confidence intervals of the medians or other quantiles within groups. Modifications of these procedures are required when censored observations are present in the data. When we compare groups of subjects, it is always good to begin with a graphical display of the data in each group. Among the various non-parametric tests one can find in the statistical literature, the Mantel-Haenzel (1959) test, currently called the "logrank" test will be used. Nowadays, the Kaplan-Meier method for estimating survival curves and the log-rank test for comparing two estimated survival curves are the most frequently used statistical tools in medical reports on survival data [30].

Log-rank test

The log rank test, developed by Mantel and Haenszel, is a non-parametric test for comparing two or more independent survival curves. Since it is a non-parametric test, no assumption about the distributional form of the data is required. This test is however

most powerful when used for non-overlapping survival curves. This test can be generalized to accommodate other tests that are equally used sometime in practice such as Generalized Wilcoxon test, Tarone-Ware test, and Peto-Peto-Prentice test. Each of these tests uses different weights to adjust for censoring that is often encountered in survival data. For instance, the Wilcoxon test weights the j^{th} failure time by n_j (the number still at risk), the Tarone–Ware test weights the j^{th} failure time by $\sqrt{n_j}$ and the PetoPeto-Prentice test weights the j^{th} failure time by the survival estimate $\tilde{S}(t_j)$ calculated over all groups combined [40] and [30]. The log rank test statistic for comparing two groups is given by:

$$Q = \frac{\left[\sum_{i=1}^{m} w_i (d_{1i} - \hat{e}_{1i})\right]^2}{\sum w_i^2 \hat{v}_{1i}}$$

Where:-

- m is the number of rank ordered events (death) times
- d_{1i} is the ordered number of events(death group 1 at event time t_i).
- $\hat{e}_{1i} = \frac{n_{1i} - d_i}{n_i}$ is the expected number of evets (death) corresponding to d_{1i}.
- n_{1i} is the number of individuals at risk in group1 just prior to event (death) time t_i.
- n_{2i} is the number of indiviuals at risk in group 2 just prior to event (deth) time t_i.
- $\hat{v}_{1i} = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$ Is the variance of the nfumber of events d1 at time t_i.
- n_i and d_i are the number of individuals at risk and number of deah in both groups (i.e., group1 and group 2) just prior to time t_i respectively.

Under the null hypothesis that two survival functions are equal, the log rank test statistic Q has an approximation of chi-square distribution with one degree of freedom ($\chi_{(1)}^2$) for large samples. The null hypothesis of equality of survival functions will be rejected for large values of Q. The most frequently used test is based on weights equal to one $w_i = 1$

. Note that the log-rank test can be extended for comparing three or more groups of survival experience.

3.2.4. Modeling Survival Data

The non-parametric methods can be useful in the analysis of a single sample of survival data, or in the comparison of two or more groups of survival times. However, in most medical studies that gives rise to survival data, supplementary information will also be recorded on each individual [10]. In this particular study, some socio-economic, demographic, and health factors may all have an impact on the time that the patients survives. Accordingly, the values of these factors(variables), which are reffered to as explanatory variables, have been recorded at the onset of the study. In order to explore the relationship between the survival experience of a patient and explanatory variables, an approach based on statistical modelling have been used.

Through a modellng approach to the analysis of survival data, we can explore how the survival experience of agroup of patients depends on the values of one or more explanatory variables, whose values have been recorded for each patient at the origin [10]. In this study, one of the objectives is to determine which of the given explanatory variables have an impact on the survival time of the patients.

3.2.4.1.Modelling the Hazard Function

In the analysis of survival data, interest centres on the risk or hazard of death at any time after the time origion of the study. As a consequence, the hazard function is modelled directly in survival analysis. The resulting models are somewhat different in form from liear models encountered in regression analysis and in the analysis of data from designed experiment, where the dependence of the mean response, or some function of it, on certain explanatory variables is modelled. However, many of the principles and procedures used in linear modelling cary over to the modelling of survival data [10].

Objectives of modelling survival data

- To determine which combiation of potential explanatory variables affect the form of the hazard function.

- To obtain an estimate of the hazard function itself for an individual so that the survivor function can be found.

This will in turn lead to an estimate of the quantities such as the median survival time, which will be a function of the explanatory variables in the model. The median survival survival time could then be estimated for the current or future patients with particular values of the explanatory variables. The resulting estimate could be particularly useful in counselling the patient about their prognosis [10].

The Cox Proportional Hazards Model

The Cox Proportional Hazard (PH) Model is a multiple regression method and is used to evaluate the effect of multiple covariates on the survival [13]. Cox (1972) proposed a semiparametric model for the hazard function that allows the addition of covariates, while keeping the baseline hazards unspecified and can take only positive values and it is defined as

$$h(t, X, \beta) = h_0(t)e^{\beta' X}, \qquad [3.13]$$

where $h(t, X, \beta)$ is the hazard function at time t with covariates $X^{|} = (X_1, X_2, ...X_P)$.

$h_0(t)$ is the arbitrary baseline hazard function that characterizes how the hazard function changes as a function of survival time.

$\beta^{|} = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of p regression parameters associated with explanatory variables.

$e^{\beta' X}$ characterizes how the hazard function changes as a function of subject covariates.

T is the failure time.

Each individual has its own hazard function of survival time. Then, the above model becomes

$$h(t, X, \beta) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}), i = 1,2,..., n \qquad [3.14]$$

where: n is total number of observations in the study.

$x_i^{|} = (x_{i1}, x_{i2}, ..., x_{ip})$ is a vector of measured covariates for the i^{th} individual (patient) which are assumed to affect the survival probability.

The Cox proportional hazard model is popular because it allows a flexible choice of covariates: time varying, time-independent, continuous and discrete. Two other issues that makes it popular are that it does not make any assumption about the underlying survival distribution and also does not require estimation of the baseline hazard rate, $h_0(t)$ to estimate the regression parameters.

Assumptions of Cox Proportional Hazards Model

- The baseline hazard function $h_0(t)$ depends on t , but not on covariates $x_1, x_2, ..., x_P$
- The hazard ratio, e^{β} , depends on the covariates $X^{|} = (X_1, X_2, ..., X_P)$ not on time.
- The covariates X_i are time-independent.

Because of the 2nd assumption it is called a proportional hazards model. To show this issue mathematically, consider two distinct values of a continuous covariate X, say, x_{i1} and x_{i2}

$h(t, X, \beta) = h_0(t)e^{\beta' X}$, Then the hazard ratio becomes

$$\frac{h(t, x_1, \beta)}{h(t, x_2, \beta)} = \frac{h_0(t)e^{\beta_i x_{i1}}}{h_0(t)e^{\beta_i x_{i2}}} = e^{\beta_i (x_{i1} - x_{i2})} \qquad [3.15]$$

which is clearly independent of time.

This reveals that the ratio of the hazard functions for two individuals with different covariate values does not vary with time.

Fitting the Cox Proportional Hazards Model

The data in survival analysis based on the sample size n are denoted by the triplet (t_i, δ_i, x_i), $i = 1,2,..., n$ where t_i is the time at which the i^{th} individual dies from the disease of interest, δ_i is the event indicator δ_i =1 if the event has occurred and δ_i =0 if it is censored (the lifetime may be right, left or interval censored), and x_i is the vector of covariates or the risk factors for the i^{th} individual.

The Cox model will be fitted by estimating the unknown regression coefficients through the maximum likelihood method. The actual likelihood function is constructed by considering the contribution of the probability that a subject with covariate value

x dies from the disease of interest at time t (i.e., $f(t, \beta, X)$), and the probability that a subject with covariate value x survives at least t time units (i .e., $S(t, \beta, X)$). That is, under the assumption of independent observations, the full likelihood function is obtained by multiplying the respective contributions of the observed triplets, a value of $f(t, \beta, X)$ for a noncensored observation and a value of $S(t, \beta, X)$ for censored observations.

Thus, the contribution of each triplet to the likelihood is the expression

$$[f(t, \beta, X)]^{\delta_i} \times [S(t, \beta, X)]^{1-\delta_i} \qquad [3.16]$$

Since the observations are assumed to be independent, the likelihood function is the product of the expression in [3.16] over the entire sample and is formulated as:

$$L(\beta) = \prod_{i=1}^{n} \left\{ [f(t_i, x_i, \beta)]^{\delta_i} \times [S(t_i, x_i, \beta)]^{1-\delta_i} \right\} \qquad [3.17]$$

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \left\{ [h(t_i, X_i, \beta) \times S(t_i, X_i, \beta)]^{\delta_i} \times [S(t_i, x_i, \beta)]^{1-\delta_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ [h(t_i, X_i, \beta)]^{\delta_i} \times [S(t_i, x_i, \beta)] \right\} \end{aligned} \qquad [3.18]$$

Thus the full likelihood function after some computations over equation [3.18] (See [10]) is then given by the expression:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \right]^{\delta_i} \qquad [3.19]$$

Cox (1972) proposed using an expression he called a partial likelihood function due to the fact that the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. In other words, the likelihood for the Cox model does not consider probabilities for all subjects. Let us consider a sample of n subjects and suppose a total of m failures occur, with m smaller than n, due to the presence of censoring. Let $t_1 < t_2 < ... < t_m$ be the m distinct ordered failure times observed and let $R(t_i)$ be the set of individuals at i^{th} failure time, which consists of all subjects with survival or censored times greater than or equal to the specified time [30].

The expression in equation [3.19] assumes that there are no tied times, and it is often modified to exclude terms when δ_i=0, yielding the partial likelihood function given as

$$L_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \qquad [3.20]$$

To obtain the maximized partial likelihood function with respect to the parameters of interest, β, we maximize the log partial likelihood function in equation [3.21].

$$l_p(\beta) = \sum_{i=1}^{m} \left\{ x_{(i)}\beta - \ln\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} \right] \right\} \qquad [3.21]$$

We obtain the maximum partial likelihood estimator by differentiating the right hand side of [3.21] with respect to β, setting the derivatives equal to zero and solving for the unknown parameters. This is known as the Newton-Raphson iterative method.

That is, for each derivative

$$U(\beta) = \frac{\partial l_p(\beta)}{\partial \beta} = \sum_{i=1}^{m} \left\{ x_{(i)} - \frac{\sum_{j \in R(t_{(i)})} x_j e^{x_j \beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \right\} = \sum_{i=1}^{m} \left\{ x_{(i)} - \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j \right\} = \sum_{i=1}^{m} \left\{ x_{(i)} - \bar{x}_{w_i} \right\} = 0 \qquad [3.22]$$

Where $w_{ij}(\beta) = \frac{e^{x_j \beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}}$ and $\bar{x}_{w_i} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j$

U (β) is called the score or gradient vector. The solution to the equation [3.22] is denoted by $\hat{\beta}$. The estimator of the variance of the estimator of the coefficient is obtained in the same manner as variance estimators are obtained in most maximum likelihood estimation applications. The estimator is the inverse of the negative of the second derivative of the log partial likelihood at the value of the estimator. Derivation of the expression in [3.22], will result in

$$\frac{\partial^2 l_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \left\{ \frac{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} \right] \left[\sum_{j \in R(t_{(i)})} x_j^2 e^{x_j \beta} \right] - \left[\sum_{j \in R(t_{(i)})} x_j e^{x_j \beta} \right]^2}{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} \right]^2} \right\} \qquad [3.23]$$

The expression in [3.23] shall be simplified using $w_{ij}(\beta)$ in equation [3.22] above. That is,

$$\frac{\partial^2 l_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij}\left(x_j - \bar{x}_{w_i}\right)^2 \qquad [3.24]$$

The negative of the 2nd derivative of the log partial likelihood in either [3.23] or [3.24] is known as the observed information and denoted by

$$I(\beta) = -\frac{\partial^2 l_p(\beta)}{\partial \beta^2} \qquad [3.25]$$

If we consider models that contain more than one covariate, the result in [3.25] becomes $I(\beta) = -\frac{\partial^2 l_p(\beta)}{\partial \beta \partial \beta'}$ which is known as the observed information matrix (Hessian matrix).

According to the Newton-Raphson procedure an estimate of β at the $(j+1)^{th}$ of the iterative procedure, $\hat{\beta}_{j+1}$, is $\hat{\beta}_{j+1} = \hat{\beta}_j + I^{-1}(\hat{\beta}_j)U(\hat{\beta}_j)$, $j = 0,1,2,\ldots$. As a result, the estimator of the variance of the estimated coefficient is the inverse of [3.25] evaluated at $\hat{\beta}$ and is

$$\hat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1} \qquad [3.26]$$

Model Diagnostics for Cox PH Model

Model-based inferences depend completely on the fitted statistical model. For these inferences to be valid in any sense of the word, the fitted model must provide an adequate summary of the data upon which it is based. The methods for assessment of a fitted proportional hazards model are essentially the same as for other regression models [30].

Residuals are used to investigate the lack of fit of a model and useful for examining different aspects of the model. The following residuals have been proposed for use by different authors in connection with the Cox regression model.

Cox-Snell, Martingale, and Deviance Residuals

The Cox-Snell residual for the i^{th} individual is given as $r_{c_i} = \hat{H}_i(t) = -\ln \hat{S}_i(t)$, where $\hat{H}_i(t)$ and $\hat{S}_i(t)$ are the estimated values of the cumulative hazard and survivor functions of the i^{th} individual at time t respectively. The Cox-Snell residuals will not be symmetrically distributed about zero and cannot be negative. Hence, the plot of the cummulative hazard of Cox-Snell residuals, $H(r_{c_i})$ versus Cox-Snell residuals, r_{c_i} gives a straight line with unit slope and zero intercept if the fitted model is correct [12].

The martingale residual is a slight modification of the Cox-Snell residuals and is defined as $r_{m_i} = \delta_i - r_{c_i}$ where, δ_i is the censoring indicator and r_{c_i} is Cox-Snell residual [5]. The martingale residuals take values between negative infinity and unity. They are uncorrelated and also have a skewed distribution with mean zero in large samples. In this respect they have properties similar to those possessed by residuals encountered in linear regression analysis [4], [10].

The deviance residuals are a normalized transformation of the martingale residuals. The deviance residual for the i^{th} individual is defined by:-

$$r_{D_i} = sign(r_{m_i})\left[-2\{r_{m_i} + \delta_i \log(\delta_i - r_{m_i})\}\right]^{1/2},$$

Where, the function $sign(\cdot)$ is the sign function which takes the value 1 if the martingale residual, r_{m_i} is positive and -1 if r_{m_i} is negative; and $\delta_i = 1$ for uncensored observation, $\delta_i = 0$ for censored observation. The deviance residuals also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Deviance residual can also be used like residuals from linear regression. The plot of the deviance residuals against the covariates can be obtained. Any unusual patterns may suggest features of the data that have not been adequately fitted for the model. Very large or very small values suggest that the observation may be an outlier in need of special attention. In a fitted Cox PH model, the hazard of death for the i^{th} individual at any time depends on the value of $\exp(\beta' x_i)$ which is called the risk score. A plot of the deviance residuals versus the risk score is a helpful diagnostic to assess a given individual on the model. Potential outliers will have deviance residuals whose absolute values are very

large. This plot will give the information about the characteristic of observations that are not well fitted by the model [70].

Schoenfeld Residuals

This overcomes the problem that the above three residuals depend heavily on observed survival time and cumulative hazard function. They are computed for each individual and covariate. It follows that, the Schoenfeld residual for the i^{th} individual and k^{th} covariate is defined as:

$$\hat{s}_{ik} = \delta_i \left[x_{ik} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} \exp\left(\hat{\beta}' X_j\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\hat{\beta}' X_j\right)} \right] \qquad [3.27]$$

where, X_j is a vector of p fixed covariates for the j^{th} individual, x_{jk} is the value of k^{th} covariate on the j^{th} individual.

Because of that, Schoenfeld residuals are defined only for the uncensored observations in which case

$\hat{s}_{ik} = x_{ik} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} \exp\left(\hat{\beta}' X_j\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\hat{\beta}' X_j\right)}$ and for each covariate it must sum to zero. In addition, they are uncorrelated and with expected value zero [58].

Most of the model diagnostics in survival data are based on the residuals stated above. The following are model diagnostics that are required to assess the model adequacy in this particular study.

Testing for the Nonlinearity of Covariates

After identifying a particular set of explanatory variables on which the hazard function depends, it is desirable to check whether the correct functional form has been adopted for the continuous covariates. The Martingale residuals can be plotted against covariates to detect nonlinearity. Nonlinearity is not an issue for categorical variables, so we only examine plots of martingale residuals against continuous covariate. LOESS

smoothed curve can be superimposed on the scatter plots to give interpretation. If the functional form observed in using the plots has some pattern, which is non linear, the covariate can be so transformed and the martingale residuals again should be plotted against the transformed covariate. A horizontal straight line which is drawn as a reference through zero would then confirm that the appropriate transformation has been used to the covariate. In addition, if the resulting smooth plot is a straight line compared to the reference line, then it shows linearity [10].

Examining Influential Observations

Another important aspect of model evaluation is through diagnostic statistics in order to identify which subjects have an unusual configuration of covariates or observations that have influence on the estimates of the parameters or on the fit of the model. In other words a fitted model is particularly sensitive to one or more observations in the data set. Such observations can be termed as influential observations. Conclusions from survival analyses are often framed in terms of estimates of the relative hazard, which depends on the estimated values of the coefficients in the Coxregression model. Thus, it is desirable to examine the influence of each observation on these estimates. The interest is about observations that influence estimate of hazard functions and the complete estimate of the model and identifications of these observations. This could be done by fitting the model to all n observations in the data set, and then fitting the same model to the sets of n-1observations obtained by omitting each of the n observations in turn. The interest is to determine if the result would change when a particular observation is removed from the analysis [10].

Suppose that $l_p(\beta)$ in equation [3.21] is log partial likelihood and $\hat{\beta}_j$ is the corresponding j^{th} parameter estimate of the model containing all the n observations and $l_p(\beta_{-i})$ be the log partial likelihood and $\hat{\beta}_{j(-i)}$ is the j^{th} parameter estimate of the model containing only the $n-1$observations after deleting the i^{th} observation, respectively. Then, the statistic $\Delta_i\hat{\beta}_j = \hat{\beta}_j - \hat{\beta}_{j(-i)}$, which is known as DFBETA, can be used as a measure of how the j^{th} parameter estimate would change if the i^{th} observation was deleted from the data set. On the other hand, the statistic,

$LD_i = 2l_p(\beta) - 2l_p(\beta_{-i})$, which is called the likelihood displacement statistic, can be used as a measure of how the maximized partial log likelihood changes if the i^{th} observation was deleted from the data set. Observations that influence a particular parameter estimate have a large absolute value of DFBETA than other observations in the data set. Observations that do influence the overall fit of the model are those which have large values of likelihood displacement statistics than the other observations in the data set [10].

Checking Cox Proportional Hazard Assumption

In order to use the Cox model, it has to be checked that the assumption of whether the effects of covariates on hazard ratio remain constant over time. This is a vital assumption of proportional hazards model and must be assessed for each covariate. Several procedures of graphical techniques and tests are proposed to investigate the proportionality assumptions in fitting the Cox model [13]. The Schoenfeld residuals are employed to assess this assumption.

The Schoenfeld residuals graphical technique can be used to assess Cox model proportionality assumption. The technique is based on individual contributions to the logpartial likelihood and measures the difference between the covariate for the i^{th} individual and a weighted average of the covariate over the risk set at each event. To check the proportionality assumption for each covariate, we plot the scaled Schoenfeld residuals against log of survival time. If the proportional hazards assumption is satisfied, the distribution of residuals over time is random, i.e., it does not show a particular trend, and the smoothed plot called Locally Weighted scatterplot smoothing (LOWESS) line summarizing the residuals should be a straight line and close to the horizontal reference line. Otherwise, a plot of scaled Schoenfeld residuals for a given covariate may reveal a violation of the proportional hazards assumption [58].

Formal tests need to detect any time dependency in particular covariates, after allowing for the effects of explanatory variables that are known. Testing the dependency of covariates on time is equivalent to testing for a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption. The

Grambsch-Therneau test of non-proportionality uses partial residuals for the test of proportional hazards assumption. In order to use this test for the i^{th} covariate, Grambsch and Therneau (1994) propose a time-varying coefficient as $\beta_i(t) = \beta_i + \gamma_i g_i(t)$. Where $\beta_i(t)$ is time varying coefficient, β_i is constant, and $g_i(t)$ is some specified function of time, usually $g_i(t) = \ln(t)$; Then, the Cox proportional hazard model for time varying coefficient with $g_i(t) = \ln(t)$ is defined as

$h(t, x_i, \beta_i(t)) = h_0(t)\exp(\beta_i(t)x)$, by substituting for $\beta_i(t)$ and $g_i(t)$ becomes

$$\begin{aligned} &= h_0(t)\exp((\beta_i + \gamma_i \ln t)x) \\ &= h_0(t)\exp(\beta_i x + \gamma_i(\ln t)x) \end{aligned} \qquad [3.28]$$

Equation [3.28] is the proportional hazards model with the interaction term, $x\ln(t)$ and main effect x_i. To test the significance of the interaction term $x\ln(t)$, we perform the test: $H_0 : \gamma = 0, versus, H_1 : \gamma \neq 0$ and we use likelihood-based tests like Wald test. If γ= 0 is not rejected, $\beta_i's$ are not time varying coefficients and hence the proportional hazards assumption is satisfied. If γ= 0 is rejected then the proportional hazards assumption is not satisfied, that leads to the need of other methods that cope with time-dependency [58].

Strategies for Analyzing of Non-Proportional Data

Suppose those statistical tests or other diagnostic techniques give strong evidence of non-proportionality for one or more covariates. To deal with this there are two popular methods: stratified Cox model and Cox regression model with time-dependent covariate which are particularly simple and can be done using available statistical software. Another alternative to consider is to use a different model. A parametric model such as an AFT model, which we will describe in next section, is more appropriate. And hence, this alternative will be considerd in such cases in this particular thesis.

3.2.4.2.Parametric Regression Models for Survival Data

The rationale for using either nonparametric or semiparametric models, in particular the semiparametric proportional hazards regression model, is to avoid having to specify the hazard function completely. The utility of the proportional hazards model stems from the

fact that a reduced set of assumptions is needed to provide the hazard ratios formed from the coefficients that are easily interpreted and clinically meaningful. However, there may be settings in which the distribution of survival time, through previous research, has a known parametric form that justi fies use of a fully parametric model to better address the goals of the analysis. These models have some advantages. In particular,

- full maximum likelihood may be used to estimate the parameters.
- the coefficients can be clinically meaningful and, for some models, are re lated to those from a proportional hazards model.
- fitted values from the model can provide estimates of survival time.
- residuals can be computed that are differences between observed and predicted values of time.

The result is that an analysis using a fully parametric model can have the look and feel of a normal errors linear regression analysis [30].

Survival time models that can be linearized by taking logs are called accelerated failure time models. The reason for this terminology is that the effect of the covariate is multiplicative on the time scale. That is, the effect of the covariate is associated with either "accelerated" or “decelerated” failure time. Where as, in the proportional hazards model, the effect of the covariates is multiplicative on the hazard scale [30].

The Gompertz PH Model

The survival and hazard function of the Gompertz distribution are given by:-

$S(t) = \exp\left(\frac{\lambda}{\theta}\left(1 - e^{\theta t}\right)\right)$ and $h(t) = \lambda \exp(\theta t)$ respectively

For $0 \leq t < \infty$ and $\lambda > 0$. The parameter θ determines the shape of the hazard function. When $\theta = 0$, the survival time then has an exponential distribution, i.e., the exponential distribution is also a special case of the Gompertz distribution. Like the Weibull hazard function, the Gompertz hazard increases or decreases monotonically. For the Gompertz distribution $\log(h(t))$ is linear with t.

Under the Gompertz PH model, the hazard function of a particular patient is given by

$h(t|X) = \lambda \exp(\theta t) \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = \lambda \exp(\beta' X) \exp(\theta t).$

It is straightforward to see that the Gompertz distribution has the PH property. But the Gompertz PH model is rarely used in practice.

Most computer software for fitting the exponential and Weibull models uses a different form of the model, AFT model, which we will describe it in the next section.

The Accelerated Failure Time Models

Although the PH model finds widespread applicability in the analysis of survival data, there are relatively few probability distributions for the survival times that can be used with this model. A model that encompasses a wide range of survival time distributions is the AFT model. In circumstances where the PH assumption is not tenable, models based on this general family may prove to be fruitful. Again, The Weibull distribution which includes exponential distribution as a special case may be adopted for distribution of survival times in AFT models, but some other probability distributions are also available [10].

Parametric AFT models are unified by the adoption of a log-linear representation of the model. This representation shows that the AFT model for survival data is closely related to the general linear model used in regression analysis. Moreover, this form of the model is adopted by most computer software packages for AFT modeling [10].

The AFT model for survival time assumes that the relationship of logarithm of survival time T_i, associated with the life time of the i^{th} individual in a study and the corresponding covariates is linear and can be written as

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\varepsilon_i \qquad [3.30]$$

In this model, $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the unknown coefficients of the values of p explanatory variables, $X_1, X_2, \ldots, X_p$, and μ, σ are two further parameters, known as the intercept and scale parameter, respectively. The quantity ε_i is a random variable used to model the deviation of the values of $\log T_i$ from the linear part of the model, and ε_i is assumed to have a particular probability distribution. In this formulation of the model, the α-parameters reflect the effect that each explanatory variable has on the survival times;

positive values suggest that the survival time increases with the values of the explanatory variable, and vice versa.

The general form of the survivor and the corresponding hazard function for the i^{th} individual in an AFT model to situations where the values of p explanatory variables have been recorded for each individual in the study can be derived from the log-linear formulation in equation [3.30]. And it is given by

$$S_i(t) = P(T_i \geq t) = P\{\exp(\mu + \alpha' x_i + \sigma\varepsilon_i) \geq t\},$$

Where $\alpha' x_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}$

Now, $S_i(t)$ can be written in the form

$$S_i(t) = P\{\exp(\mu + \sigma\varepsilon_i) \geq {}^{t}\!/_{\exp(\alpha' x_i)}\},$$

and the baseline survivor function, $S_0(t)$, the survivor function of an individual for whom $x = 0$, is

$$S_0(t) = P\{\exp(\mu + \sigma\varepsilon_i) \geq t\}$$

It then follow that

$$S_i(t) = S_0\left({}^{t}\!/_{\exp(\alpha' x_i)}\right), \qquad [3.31]$$

Equation [3.31] is the general form of the survivor function for the i^{th} individual in an AFT model; the acceleration factor is $\exp(-\alpha' x_i)$ for the i^{th} individual. The corresponding relationship between the hazard functions is obtained by taking natural logarithms of sides of equation [3.31], multiplying by -1, and differentiating with respect to t, leads to

$$h_i(t) = \exp(-\alpha' x_i) h_0\{{}^{t}\!/_{\exp(\alpha' x_i)}\}, \qquad [3.32]$$

A general expression for the p^{th} percentile of the distribution of survival times in AFT models follows from the result that an AFT model can be derived from many probability distributions for ε_i, although some are more tractable than others.

This can be shown as follows

$S_i(t) = P(T_i \geq t) = P(\log T_i \geq \log t)$.

From equation [3.30],

$$S_i(t) = P(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\varepsilon_i \geq \log t),$$

$$= P\left(\varepsilon_i \geq \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right) \qquad [3.33]$$

If we now write $S_{\varepsilon_i}(\varepsilon)$ for the survivor function of the random variable ε_i in the log-linear model of equation [3.30], the survivor function of the i^{th} individual can, form equation [3.33], be expressed as

$$S_i(t) = S_{\varepsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right). \qquad [3.34]$$

Equation [3.34] shows how the survivor function for T_i can be found from the survivor function of the distribution of ε_i. Then the p^{th} percentile for the i^{th} individual, $t_i(p)$, is given by

$S_i\{t_i(p)\} = \frac{100 - p}{100}$, and using equation [3.33],

$$P\left(\varepsilon_i \geq \frac{\log t_i(p) - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right) = \frac{100 - p}{100}.$$

If $\varepsilon_i(p)$ is used to denote the p^{th} percentile of the distribution of ε_i, then

$$S_{\varepsilon_i}\{\varepsilon_i(p)\} = p\{\varepsilon_i \geq \varepsilon_i(p)\} = \frac{100 - p}{100}.$$

consequently,

$$\varepsilon_i(p) = \frac{\log t_i(p) - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma},$$

and so

$$t_i(p) = \exp\{\sigma\varepsilon_i(p) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}\} \qquad [3.35]$$

equation [3.35] is the general expression for the p^{th} percentile for the i^{th} individual, $t_i(p)$ in AFT model. It can be written in the form

$$t_i(p) = \exp(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}) t_0(p), \qquad [3.36]$$

Where $t_0(p)$ is p^{th} percentile for a baseline individual for whom all explanatory variables take the value zero. This confirms that the α-coefficients can be interpreted in terms of the effect of the explanatory variables on a particular percentile of survival times [10].

Particular choices for the distribution of ε_i in the log-linear formulation of the AFT model equation [3.30], leads to distributions for the random variable associated with the survival time of the i^{th} individual. Parametric AFT models based on the Weibull, log-logistic and log-normal distributions for the survival times are most commonly used in practice.

The Weibull AFT Model

The Weibull distribution which includes the exponential distribution as a special case can also be parameterized as an AFT model, and they are the only family of distributions to have both PH and AFT property. The results of fitting a Weibull model can therefore be interpreted in either framework. Then the Weibull distribution is very flexible model for time-to-event data. It has a hazard rate which is monotonically increasing or decreasing and constant when the shape parameter $\gamma = 1$. If $T_i = \exp(\mu + \alpha' x_i + \sigma\varepsilon_i)$ has a Weibull distribution In terms of the log-linear representation of the model in equation [3.30], then ε_i does in fact have a type of extreme value distribution known as Gumbel distribution. This is an asymmetric distribution with survivor function given by:

$$S_{\varepsilon_i}(\varepsilon) = \exp(-e^{\varepsilon}), \text{for} -\infty < \varepsilon < \infty \qquad [3.37]$$

Then form equation [3.34], the survivor function of T_i is given by:

$$S_i(t) = \exp\left\{-\exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right)\right\}. \qquad [3.38]$$

This can be expressed in the form

$$S_i(t) = \exp\left(-\lambda_i t^{1/\sigma}\right),$$

Where:-

$\lambda_i = \exp\left\{-\frac{(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi})}{\sigma}\right\}$ is the scale parameter and σ^{-1} is shape parameter.

The cummulative hazard can be obtained as follows:

$$H_i(t) = -\log S_i(t) = \exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right),$$

Which can also be expressed as $\lambda_i t^{1/\sigma}$ and the hazard function is obtained by defferentiating the cummulative hazard with respect to t and hence it is given as follows:

$$h_i(t) = \frac{1}{\sigma t}\exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right), \qquad [3.39]$$

Or $h_i(t) = \lambda_i \sigma^{-1} t^{\sigma^{-1}-1}$.

We now reconcile this form of the model with that for the Weibull PH model. From equation [3.29], above the survivor function for the i^{th} individual is

$$S_i(t) = \exp\left\{-\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\right)\lambda t^{\gamma}\right\}, \qquad [3.40]$$

In which λ and γ are parameters of the Weibull baseline hazard function. There is a direct correspondence between equation [3.38] and [3.40] in the sense that

$$\lambda = \exp\left(-\mu/\sigma\right), \quad \gamma = \sigma^{-1}, \quad \beta_j = -\alpha_j/\sigma, \text{ for } j = 1, 2, \ldots, p$$

We therefore deduce that the log-linear model in which ε_i has Gumbel distribution, provides an alternative representation of the Weibull PH model.

In this form of the model, the p^{th} percentile of the survival time distribution for the i^{th} individual is the value $t_i(p)$, which is such that $S_i\{t_i(p)\} = 1 - \left(p/100\right)$, where $S_i(t)$ is as given in equation [3.38] straightforward algebra leads to the result that

$$t_i(p) = \exp\left[\sigma \log\left\{-\log\left(\frac{100 - p}{100}\right)\right\} + \mu + \alpha' x_i\right] \qquad [3.41]$$

The Log-Logistic AFT Model

The log-logistic distribution provides the most commonly used AFT model. Unlike the Weibull distribution, it can exhibit a non-monotonic hazard function which increases at early times and decreases at later times. It is similar in shape to the log-normal distribution but its cumulative distribution function has a simple closed form, which becomes important computationally when fitting data with censoring [36].

When the survival times have log-logistic distribution with parameters θ, k, then the baseline hazard function is given by:

$$h_0(t) = \frac{e^{\theta} k t^{k-1}}{1 + e^{\theta} t^{k}} \qquad [3.42]$$

And the corresponding survivor function is

$$S_0(t) = \frac{1}{1 + e^{\theta} t^{k}} \qquad [3.43]$$

The log-linear form of the accelerated time model in equation [3.30] also provides a representation of the log-logistic distribution. Suppose that in this formulation, ε_i now has a logistic distribution with zero mean and variance $\pi^2/3$, so that the survivor function of ε_i is $S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + e^{\varepsilon}}$.

Using equation [3.34], the survivor function of T_i is then

$$S_i(t) = \left\{1 + \exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right)\right\}^{-1} \qquad [3.44]$$

From equation [3.31] and [3.43], the survivor function of T_i, is derived as follows:

$$S_i(t) = \frac{1}{1 + e^{\theta - k\eta_i} t^{k}}. \qquad [3.45]$$

It then follows that T_i has a log-logistic distribution with parameters $\theta - k\eta_i$, k. where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}$. The log-logistic distribution therefore has the AFT property. However, this distribution doesn't have the PH property [10].

On comparing expression [3.45] with that for the survivor function in equation [3.44], we see that the parameters θ and k can be expressed in terms of μ and σ. Specifically,

$$\theta = -\mu/\sigma, \; k = 1/\sigma \qquad [3.46]$$

Where with k>0; When $k \leq 1$, the hazard rate decreases monotonically and when k>1, it increases from zero to a maximum and then decreases to zero.

The hazard function for the i^{th} individual is obtained by applying log to the survivor function in equation [3.44] and multiplying by -1 then differentiating the result with respect to t gives

$$h_i(t) = \frac{1}{\sigma t}\left\{1 + \exp\left[-\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right)\right]\right\}^{-1} \qquad [3.47]$$

And the p^{th} percentile of the survival time distribution is from equation [3.44], or the general result in equation [3.35], is

$$t_i(p) = \exp\left\{\sigma \log\left(\frac{p}{100 - p}\right) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}\right\}. \qquad [3.48]$$

The median survival time can be obtained by substituting $p = 50$ in equation [3.48].

Note that, the natural logarithm of odds ratio in log-logistic AFT model is independent of time. Therefore, the log-logistic regression model is a proportional odds model, not a proportional hazards model. i.e., the log-logistic model is the only parametric model with both a proportional odds and an accelerated failure-time representation [35]. And hence using equation [3.43], the baseline odds of survival beyond t is given by $\frac{S_0(t)}{1 - S_0(t)} = \frac{1}{e^{\theta} t^k}$,

Similarly, using equation [3.44], the odds of survival beyond t for the i^{th} individual is

$$\frac{S_i(t)}{1 - S_i(t)} = \frac{1}{\exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}}{\sigma}\right)} = \frac{1}{\lambda t^{\gamma} \exp(\beta' x_i)}$$

$$= \frac{1}{\lambda t^{\gamma}} \exp(-\beta' x_i) = \exp(-\beta' x_i)\left[\frac{S_0(t)}{1 - S_0(t)}\right]$$

$$\Rightarrow \frac{S_i(t)}{1 - S_i(t)} = \exp(-\beta' x_i)\left[\frac{S_0(t)}{1 - S_0(t)}\right] \qquad [3.49]$$

Where $e^{\theta} = \exp\left(-\mu/\sigma\right) = \lambda$, $k = 1/\sigma = \gamma$, and $\beta_j = -\alpha_j/\sigma$

Therefore, from equation [3.49], we can see that the factor $\exp(-\beta' x_i)$ is an estimate of how much the baseline odds of survival at any time changes when an individual has

covariate x_i. Moreover, to check whether the log-logistic distribution is tenable to fit a particular survival data set, a plot of $\log\left[\frac{S_i(t)}{1 - S_i(t)}\right]$ versus $\log t$ is used and it should be linear if the log-logistic distribution is appropriate.

The Log-Normal AFT Model

If the survival times are assumed to have a log-normal distribution, the baseline survival function is given by $S_0(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$,

Where μ and σ are unknown parameters. Under AFT model, the survivor function for the i^{th} individual, is then $S_i(t) = S_0\left(e^{-\eta_i} t\right)$,

Where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi}$ is a linear combination of the values of p explanatory variables for the i^{th} individual. Therefore,

$$S_i(t) = 1 - \Phi\left(\frac{\log t - \eta_i - \mu}{\sigma}\right), \qquad [3.50]$$

which is the survivor function of an individual whose survival times have a lognormal distribution with parameters $\mu + \eta_i$ and σ. The lognormal distribution therefore has the AFT property.

In the log-linear formulation of the model, the random variable associated with the survival time of the i^{th} individual has a lognormal distribution if $\log T_i$ is normally distributed. We therefore take ε_i in equation [3.30] to have a standard normal distribution, so that the survivor function of ε_i is $S_{\varepsilon_i}(\varepsilon) = 1 - \Phi(\varepsilon)$.

The cumulative hazard, and hazard function, of ε_i are

$H_{\varepsilon_i}(\varepsilon) = -\log\{1 - \Phi(\varepsilon)\}$, and $h_{\varepsilon_i}(\varepsilon) = \frac{f_{\varepsilon_i}(\varepsilon)}{S_{\varepsilon_i}(\varepsilon)}$, respectively

Where $f_{\varepsilon_i}(\varepsilon)$ is the density function of a standard normal random variable, given by

$$f_{\varepsilon_i}(\varepsilon) = \frac{1}{\sqrt{(2\pi)}} \exp\left(-\varepsilon^2/2\right).$$

The random variable T_i, in AFT model, then has a lognormal distribution with parameters $\mu + \alpha' x_i$ and σ. The survivor function of T_i is as given in equation [3.50] and the hazard function can be obtained from the general form in equation [3.32].

The p^{th} percentile of the distribution of T_i, from equation [3.35], is

$$t_i(p) = \exp\left\{\sigma\Phi^{-1}\left(p/100\right) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi}\right\},$$

And in particular, $t(50) = \exp(\mu + \alpha' x_i)$ is the median survival time for the i^{th} individual.

The Generalized Gamma AFT Model

In this section we introduce a regression model that is based on the general form of the gamma distribution. Therefore, in this study whenever we speak off gamma AFT model, we mean that the generalized gamma AFT model since it includes relatively a wide range of family distribution as its special case [35]. Let's first define some functions which may be used in the rest of this section.

The gamma function is a well-known function which is defined as follows:

$$\begin{aligned}\Gamma(\gamma) &= \int_0^\infty x^{\gamma-1}e^{-x}dx \\ &= (\gamma - 1)! \quad \text{when } \gamma \text{ is a positive integer}\end{aligned} \qquad [3.51]$$

Another useful function to know in this section is known as the incomplete gamma function and it is mathematically defined as:

$$I(s, \gamma) = \begin{cases}\dfrac{1}{\Gamma(\gamma)}\int_0^s u^{\gamma-1}e^{-u}du & \text{if } s \geq 0 \\ 0 & \text{if } s < 0\end{cases} \qquad [3.52]$$

Then, the three-parameter Generalized Gamma survivor and density functions respectively are

$$S(t) = \begin{cases}1 - I(\gamma, u), & \text{if } k > 0 \\ 1 - \Phi(z), & \text{if } k = 0 \\ I(\gamma, u), & \text{if } k < 0\end{cases} \qquad [3.53]$$

$$f(t) = \begin{cases} \frac{\gamma^{\gamma}}{\sigma t\sqrt{\gamma}\Gamma(\gamma)} \exp\left(z\sqrt{\gamma} - u\right), & \text{if } k \neq 0 \\ \frac{1}{\sigma t\sqrt{2\pi}} \exp\left(-z^{2}/2\right), & \text{if } k \neq 0 \end{cases} \quad [3.54]$$

Where $\gamma = |k|^{-2}$, $z = sign(k)\{\log(t) - \mu\}/\sigma$, $u = \gamma \exp(|k|z)$, $\Phi(z)$ is the standard normal cumulative distribution function, $I(a, x)$ is incomplete gamma function, and $sign(k)$ is a mathematical function which returns the sign of k: -1 if k < 0, 0 if k= 0, 1 if k > 0, and missing if k is missing. In STATA, this model is implemented by parameterizing $\mu_j = X_j\beta$ and treating the parameters k and σ as ancillary parameters to be estimated from the data.

The hazard function of the generalized gamma distribution is extremely flexible, allowing for many possible shapes, including as special cases the Weibull distribution when $k = 1$, the exponential when $k = 1$ and $\sigma = 1$, and the lognormal distribution when $k = 0$. The generalized gamma model is, therefore, commonly used for evaluating and selecting an appropriate parametric model for the data. The Wald or likelihood-ratio test can be used to test the hypotheses that $k = 1$ or that $k = 0$.

The hazard function can be found by applying log to the survivor function; multiplying by -1 then differentiating the result with respect to t. moreover, the p^{th} percentile of the survival time distribution can be obtained by applying equation [3.36]. Furthermore, the median survival time ratio for the j^{th} individual can be obtained by dividing the median survival time of that individual by the baseline median survival time and it is given by

$$TR_i(50) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\right) \quad [3.55]$$

Fitting Parametric Models

AFT models are fitted using the method of maximum likelihood. The likelihood function is best derived from the log-linear representation of the model, after which iterative methods are used to obtain the estimates. The likelihood of the n observed survival times, $t_1, t_2, ..., t_n$, is form equation [3.17] given by

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \left\{ [f_i(t_i)]^{\delta_i} \times [S_i(t_i)]^{1-\delta_i} \right\},$$

Where $f_i(t_i)$ and $S_i(t_i)$ are the density and survivor functions for the i^{th} individual at t_i, and δ_i is the event indicator for the i^{th} observation, so that δ_i is unity if the i^{th} is an event and zero if it is censored. Now, from equation [3.34], $S_i(t_i) = S_{\varepsilon_i}(z_i)$,

Where $z_i = \left(\log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}\right) / \sigma$, and differentiation with respect to t gives $f_i(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}(z_i)$

The likelihood function can then be expressed in terms of the survivor and density function of ε_i, giving

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} [\sigma t_i]^{-\delta_i} \left[f_{\varepsilon_i}(z_i)\right]^{\delta_i} \left[S_{\varepsilon_i}(z_i)\right]^{1-\delta_i} .$$

The log-likelihood function is then

$$\log L(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \left\{ -\delta_i \log(\sigma t_i) + \delta_i \log f_{\varepsilon_i}(z_i) + (1 - \delta_i) \log S_{\varepsilon_i}(z_i) \right\},$$

And the maximum likelihood estimates of the $p + 2$ unknown parameter, μ, σ and $\alpha_1, \alpha_2, \ldots, \alpha_p$, are found by maximizing this function using the Newton-Raphson procedures.

Checking the Adequacy of Parametric Models

The graphical methods can be used to check if a parametric distribution fits the observed data. Specifically, if the survival time follows a Weibull distribution, a plot of $\log[-\log S(t)]$ versus $\log t$ should yield a straight line with slope of 1. If the plots are parallel but not straight, then PH assumption holds but not the Weibull. If the lines for two groups are straight but not parallel, the Weibull assumption is supported but the PH and AFT assumptions are violated. The log-logistic assumption can be graphically evaluated by plotting $\log[(1 - S(t))/S(t)]$ versus $\log t$ .If the distribution of survival function is log-logistic, then the resulting plot should be a straight line. For the log-normal distribution, a plot of $\Phi^{-1}[1 - S(t)]$ versus $\log t$ should be linear.

Using Quantile-Quantile Plot

An initial method for assessing the potential for an AFT model is to produce a quantile-quantile plot. For any value of p in the interval $(0,100)$, the p^{th} percentile is $t(p) = S^{-1}\left(\frac{100 - p}{100}\right)$. Let $t_0(p)$ and $t_1(p)$ be the p^{th} percentiles estimated from the survival functions of the two groups of survival data. The percentiles for the two groups may be expressed as

$$t_0(p) = S_0^{-1}\left(\frac{100 - p}{100}\right), \qquad t_1(p) = S_1^{-1}\left(\frac{100 - p}{100}\right),$$

Where $S_0(t)$ and $S_1(t)$ are the survival functions for the two groups. So we can get

$$S_1[t_1(p)] = S_0[t_0(p)] = \frac{1}{2}.$$

Under the AFT model, $S_1(t) = S_0(t/\eta)$, and so $S_1[t_1(p)] = S_0[t_1(p)/\eta]$.

Therefore, we get $t_0(p) = \eta^{-1} t_1(p)$.

The percentiles of the survival distributions for two groups can be estimated by the K-M estimates of the respective survival functions. A plot of percentiles of the K-M estimated survival function from one group against another should give an approximate straight line through the origin if the accelerated failure time model is appropriate. The slope of this line will be an estimate of the acceleration factor η^{-1} [34].

Using Statistical Criteria

We can use statistical tests or statistical criteria to compare all these AFT models. Nested models can be compared using the likelihood ratio test. The exponential model, the Weibull model and log-normal model are nested within gamma model. For comparing models that are not nested, the Akaike information criterion (AIC) can be used instead, which is defined as $AIC = -2l + 2(k + c)$,

Where l is the log-likelihood; k is the number of covariates in the model and c is the number of model-specific ancillary parameters. The addition of $2(k + c)$ can be thought of as a penalty if non-predictive parameters are added to the model. Lower values of the

AIC suggest a better model. But there is a difficulty in using the AIC in that there are no formal statistical tests to compare different AIC values. When two models have very similar AIC values, the choice of model may be hard and external model checking or previous results may be required to judge the relative plausibility of the models rather than relying on AIC values alone [34].

4. Results of Statistical Analysis and Discussions

4.1. Data Set Summary and Descriptive Analysis

The first step in analysis of statistical data is showing the nature of the entire data set using descriptive methods. However, meaningful conclusions could be drawn from the results of analysis on the basis of well managed data besides the reliability of the information gathered. Hence, descriptive analysis has been made in order to get some information about the distribution of survival time between/among categories of factors summarized in Table: 4.1. Therefore, potential risk factors which are expected to have significant effect on survival of patients were predicted for further analysis by descriptive analysis. The baseline values of such factors were recorded for every individual patient made eligible for the study in the accrual period.

For this particular study, there were 6200 patients under ART follow-up in the accrual period out of which only 3042 patients were eligible for the study. The remaining 3158 patients were excluded from analysis because important attributes were not recorded for them. In other words, patients included for the analysis were those for whom complete information was recorded about their baseline characteristics and their survival status were known to be dead at a certain time t or alive until the study time. Several variables were expected to affect the survival time of individual HIV/AIDS patients under ART follow-up. Some of the independent variables were basically numerical whereas others were categorical. Descriptive summary statistics associated with numerical covariates are shown in the Appendix, Table7. 1.

Similarly, Table: 4.1 shows categorical covariates in which labels of each factor were assigned numerical values so that data entry was easier. The CD4 count (percent) is numerical by nature but it was categorized in to labels because the values associated with this variable were obtained in percentage for some individuals particularly, individuals of age less than 12 and as actual count for others. Therefore categorizing CD4 percent was believed to be better way and convenient for this particular study. In this regard, actual CD4 counts were allocated to corresponding CD4 percent category (label) in such a way that a CD4% of 12-15% is the same as a count of under 200 cells/mm^3; a CD4% of 29% is the same as a count of over 500 cells/mm^3 but there is a wide range for higher values;

CD4 counts between 200 cells/$mm^3$and 500 cells/mm^3 exclusive were categorized as CD4% of 16-28% [[1]].

As it can be seen from (Table7. 1 in the Appendix), summary statistics associated with continuous covariates were not expressed in terms of patients' survival time like categorical covariates in Table: 4.1. Rather they were given to describe average age, weight and household size of HIV/AIDS patients who were under ART follow-up and considered to be eligible for this particular study. Minimum, maximum and standard deviations given for each of these covariates could be helpful to have a general insight about age, weight and household size of individual patients in this particular study. The presence of independent effect of each of these continuous covariates on survival time has been assessed in section 4.3 by Cox PH Model (Table: 4.3). The effect for age and weight appeared to be significant independently whereas household size showed statistically insignificant effect.

In the same manner, the independent effect of categorical covariates was assessed by Non-Parametric analysis as well as Cox PH model.

[1]*The immune system contains lots of different cells. The two main types of lymphocytes are T cells and B cells. CD4 cells are a type of T cells. So the CD4% looks at the CD4 count in relation to other immune cells. CD4 counts are not used for children under 12 years old, who are monitored by CD4 percentage. This is because we are born with very high CD4 levels (several thousand cells/*mm^3*).*

Table: 4.1: Description of Survival time by Categorical Covariates

Covariates	Label	Value	No. of Patients	No. of Events	Median Survival Time
Gender	Female	0	1790	271	124.7
	Male	1	1252	245	112.53
Functional Status	Working	0	2042	234	
	Ambulatory	1	815	210	111.53
	Bedridden	2	185	72	109.53
WHO Clinical Stage	I	0	129	2	
	II	1	433	23	
	III	2	1875	226	
	IV	3	605	265	103.53
CD4 Percent	12-15%	1	2460	479	112.53
	16-28%	2	549	37	
	Above28%	3	33	0	
Cotrimoxazol	Yes	0	2418	427	123.7
	No	1	624	89	124.7
TB Status	Negative	0	532	30	
	Positive	1	2510	486	123.7
Marital Status	Never Married	0	1074	151	
	Married	1	1950	365	112.53
	Widowed or Divorced	2	18	0	
Educational Status	Educated	0	1218	244	112.53
	Not Educated	1	1824	272	124.7
Opportunistic Infections	No	0	42	4	
	Yes	1	3000	512	124.7
Risk Behavior	No	0	228	6	
	Yes	1	2814	510	123.7
Occupational Status	Employed	0	2814	509	124.7
	Unemployed	1	228	7	

According to the results in Table: 4.1, differences in magnitudes of median survival time could be observed when Comparison is made between/among categories of covariates. However, it could be difficult to determine whether the observed difference is statistically significant. But it is possible to know which group of patients is likely to survive more in terms of time as compared to other group(s) defined by a particular covariate if the observed difference is assumed to be significant. Hence, Non-Parametric analysis was made in order to minimize such uncertainty in descriptive analysis of survival data and to predict possible set of covariates on which further analysis should be made.

4.2. Non-Parametric Analysis

In this part of the analysis, Survival time of individual patients was estimated for each level of categorical covariates using the K-M method described in Section 3.2.3.2 and compared using log-rank test (Section 3.2.3.3). The K-M curves for each category of study factors (categorical covariates) provide an initial insight into the shape of the survival function for levels of study factors. The log-rank test is used to compare survival time distributions among categories. The K-M estimate of the survivor function curves for Gender of patients is shown in Figure: 4.1. Similarly, K-M curves based on separate calculation of survivor functions by groups defined by each of other categorical covariates are shown in (Appendix, Figure7. 1). The corresponding log-rank test result which has been used to compare survival distribution is given in Table: 4.2.

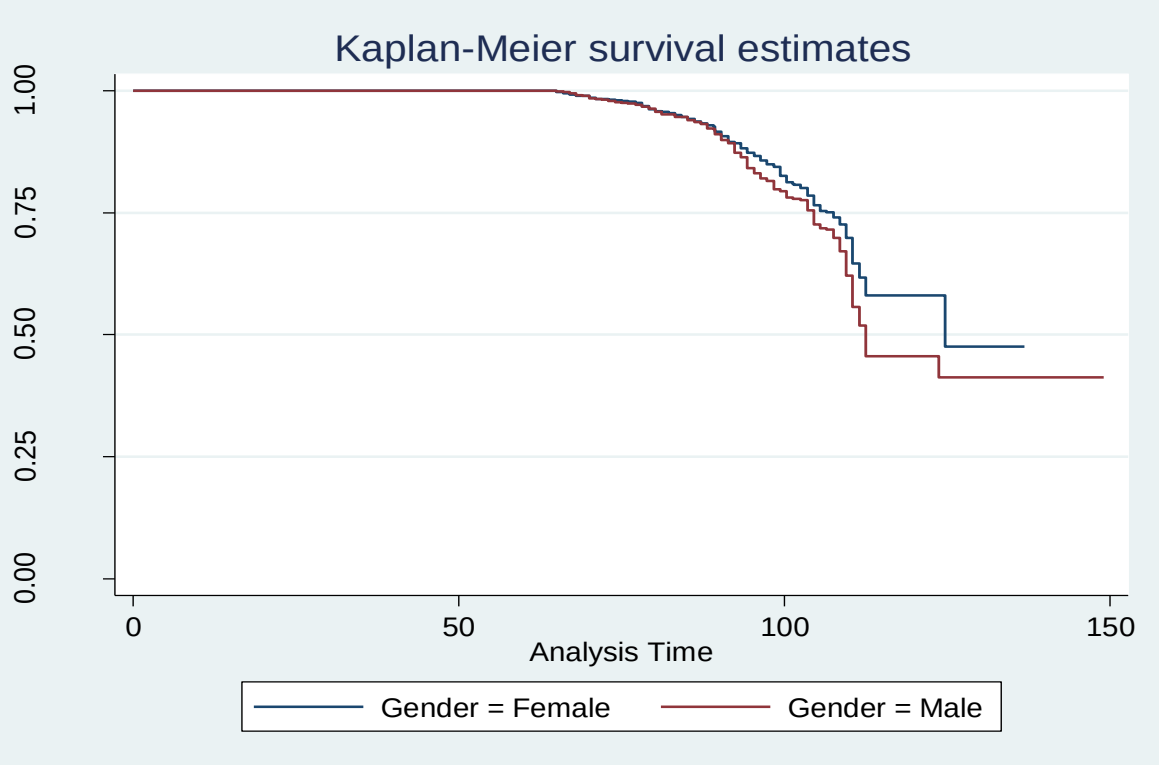


Figure: 4.1 : K-M Estimate of the Survivor Function for Categories of Gender

According to the result in Table: 4.2, by the log-rank test for equality of survivor function between Male and Female the p-value was 0.011 which is associated with high value of calculated chi-square. Therefore, there is statistically significant difference in the distribution of survival time between male and female at 5% level of significance. The

same way as it was described in Table: 4.1, Figure: **4.1** also showed that Female patients are more likely to survive longer than Male patients. However, adjusted effect of Gender appeared to be insignificant by Semi-Parametric and Parametric analysis in which the effect of a particular factor is assessed relative to all other covariates considered in the analysis (Section4.3).

Furthermore, Functional Status, WHO clinical Stage, CD4 Percent, TB Status, Educational Status, Opportunistic Infections and Risk Behavior were identified to have statistically significant effect by Non-Parametric analysis. However, adjusted effect of Educational Status was insignificant as it could be shown in the next section. The P-Values for Educational Status and Risk Behavior were 0.008 and 0.007 respectively. Whereas, Functional Status, WHO clinical Stage, CD4 Percent, TB Status and Opportunistic Infections have, P-Value=0.000 (See Table: **4.2**).

Table: 4.2: Independent Log-Rank Test for equality of survival distributions for the different levels of Categorical Covariates

Categorical Covariates	Chi-Square	df	P-value
Gender	6.548	1	.011
Functional Status	80.363	2	.000
WHO Clinical Stage	370.965	3	.000
CD4 Percent	33.724	2	.000
Cotrimoxazol	2.748	1	.097
TB Status	22.021	1	.000
Marital Status.	2.119	2	.347
Educational Status	6.929	1	.008
Occupational Status	.242	1	.623
Opportunistic Infections	13.327	1	.000
Risk Behavior	7.337	1	.007

In other words, the remaining categorical covariates such as Cotrimoxazol, Marital Status and Occupational Status which were expected to have effect on survival of individual HIV/AIDS patients at the beginning of the study were predicted to be insignificant at 5% level of significance by the Non-Parametric analysis in this particular study. The effect of continuous covariates Age, Weight and Household Size was of course not assessed yet by

the Non-Parametric analysis. However, all covariates were assessed once again by the Cox PH model independently as well as in a conditional (adjusted) manner in Section4.3.

4.3. Cox PH Model

In section 4.2, patients' survival time distribution has been compared between/among categories of each covariate independently on Non-Parametric basis. Here, the effect of such factors including continuous covariates was assessed by using Cox PH model independently as well as relatively. As predicted by the Non-Parametric Analysis, Gender, Functional Status, WHO clinical Stage, CD4 Percent, TB Status, Educational Status, Opportunistic Infections and Risk Behavior were significant in Cox PH model analysis independently. Age, Weight and Household size were also identified to have significant effect by the Cox PH model in the independent analysis (See Table: **4.3**). Cotrimoxazol, Marital Status and Occupational Status were insignificant in neither adjusted nor independent analysis.

Table: 4.3: Independent Semi-Parametric Analysis of Covariates Effect (Cox PH Model)

Covariates	Coefficients	SE	Wald	P-Value	Hazard Ratio
Age	.014	.004	10.503	.001	1.014
Gender	.223	.088	6.398	.011	1.250
Functional Status			74.849	.000	
Functional Status(1)	.666	.095	49.054	.000	1.947
Functional Status(2)	.962	.135	50.784	.000	2.616
Weight	-.016	.004	13.756	.000	.984
WHO Clinical Stage			297.854	.000	
WHO Clinical Stage(1)	.970	.737	1.729	.188	2.637
WHO Clinical Stage(2)	1.350	.710	3.613	.057	3.858
WHO Clinical Stage(3)	2.811	.710	15.680	.000	16.626
CD4Percent			25.654	.000	
CD4Percent(1)	-.865	.171	25.643	.000	.421
CD4Percent(2)	-11.135	106.067	.011	.916	.000
Cotrimoxazol	-.191	.117	2.688	.101	.826
TB Status	.868	.192	20.394	.000	2.383
Marital Status			1.209	.546	
Marital Status(1)	.106	.097	1.200	.273	1.112
Marital Status(2)	-8.937	95.857	.009	.926	.000
Education	-.229	.088	6.769	.009	.795
Occupation	-.186	.381	.237	.626	.831

Opportunistic Infections	2.138	.712	9.014	.003	8.485
Risk Behavior	1.055	.411	6.580	.010	2.872
Household Size	.070	.027	6.883	.009	1.072

Adjusted effect of covariates was then estimated by multi-factor analysis of Cox PH model after identifying potential risk factors. However, some factors which were significant in the independent analysis appeared to be insignificant in the multi-factor analysis due to relative importance.

As a result, best possible combination of significant covariates (factors) was selected to fit the desired model by using Strategy For Model Selection described in Section3.2.4.1. In this particular case, stepwise Forward Selection and Backward Elimination of covariates were performed on the basis of Wald Test in such a way that variables predicted as significant by Non-Parametric analysis had 5% chance to be included in the model and a probability of 0.1 for exclusion. Exclusion of Covariates from a model when they were significant in the independent analysis is to mean, inclusion of such factor will add no further expression of variation in survival time or function of it than already expressed by the included covariates since the presence of association among covariates might be reason for such removal. Although conditional selection of covariates is also possible based on interest or biological importance, such procedure has not been used in this particular study since the main purpose was to compare Cox PH and AFT models in determining factors related to survival time of HIV/AIDS patients. However, minimizing uncertainty associated with effects of some factors was taken as specific objective of this study. Therefore, stepwise selection of a set of significant variables that results in minimum $-2\log likelihood$ (for comparison of nested models) and AIC (for comparison of other alternative models) was performed to fit the final model. (See variable selection procedures in Section 3.2.4.1).

Therefore, after Household Size (P-Value= 0.768) and Education (P-Value= 0.5313) excluded by stepwise Backward Elimination procedure at 10% level of significance (probability of removal), WHO clinical Stage, CD4 Percent and TB Status were selected to be included in the Cox PH model at 5% level of significance as the most significant factors affecting survival of HIV/AIDS patients with P-Value=0.000. Similarly, Age (P-

Value=0.0047), Functional Status (P-Value=0.0163), Opportunistic Infections (P-Value=0.0285) and Risk Behavior (P-Value=0.0317) were also statistically significant in the Stepwise Forward Selection procedure. On the other hand, Gender and Weight were removed from the model since they were not selected by stepwise forward procedure and inclusion of both/any of these covariates couldn't results in much decrease in statistical information criterion statistics, although they were not removed by Stepwise Backward Elimination procedures. Hence, the final Cox PH model was fitted only for selected significant covariates by Stepwise Forward procedure based on results in Table: **4.4**.

Table: 4.4: Adjusted Semi-Parametric Analysis of Covariates Effect (Cox PH Model)

Covariates	Categories	Coeff.	SE	P-Value	Haz. Ratio	95% CI for Haz. Ratio	
						LL	UL
WHO Clinical Stage	I	Ref.			1		
	II	1.008	.737	.172	2.740	.65	11.623
	III	1.332	.711	.061	3.789	.94	15.263
	IV	2.684	.712	.000	14.64	3.63	59.060
CD4Percent	12-15%	Ref.			1		
	16-28%	-.660	.172	.000	.517	.37	.724
	Above28%	-10.68	105.62	.919	.000	.00	1.828E+085
TB Status	Negative	Ref.			1		
	Positive	.760	.196	.000	2.139	1.458	3.137
Functional Status	Working	Ref.			1		
	Ambulatory	.302	.099	.002	1.353	1.115	1.643
	Bedridden	.221	.143	.122	1.247	.943	1.650
Opportunistic Infections	No	Ref.			1		
	Yes	1.510	.726	.038	4.525	1.091	18.770
Risk Behavior	No	Ref.			1		
	Yes	.964	.421	.022	2.621	1.150	5.976
Age		.010	.005	.038	1.010	1.001	1.019

In Cox PH model analysis, a coefficient associated with continuous covariate is a change in the log hazard ratio due to a unit increase in the value of the variable under consideration holding the value of other covariates constant. The corresponding Hazard Ratio can be obtained by exponentiation of the coefficient and it is an adjusted multiplicative effect on hazards of death for a unit change in the covariate value.

Similarly, a coefficient associated with a particular category of a categorical covariate is adjusted additive effect on log hazard ratio for change of level from reference to that particular category of the covariate. The corresponding hazard ratio is a multiplicative effect on hazard of death for such change in levels of covariates holding the effects of other covariates constant. 95% CI for hazard ratio including the value 1 shows that statistical non-significance; 95% CI which does not include the value 1 shows statistical significance at 5% level of significance (Section3.2.4.1).

Consequently, the multiple-Covariates Cox PH model was then fitted to predict the effect of significant covariates in such a way that $k-1$ dummy variables have been created for categorical covariates having k levels so that one of the categories of each covariate was taken as a reference. The first category of each categorical covariate has been taken as a reference in this particular case. Categorical predictor WHO Clinical Stage had four levels and hence, three dummy variables ($WHOCS_{2i}$, $WHOCS_{3i}$, $WHOCS_{4i}$) have been created to include this predictor in the model taking WHO Clinical Stage I as reference group. Similarly, CD4 percent had three categories and included by ($CD4_{2i}, CD4_{3i}$) taking CD4% of 12-15% as reference; each of the categorical predictors TB Status, Opportunistic Infections and Risk Behavior had two levels and therefore, included as TB_{2i}, OI_{2i} and RB_{2i} by taking TB Status negative, Opportunistic Infections no and Risk Behavior no as reference respectively; Functional status had three categories and therefore included as (FS_{2i}, FS_{3i}) and working as reference. After a multiple covariate model of main effects has been built, then predictors were checked for all possible interactions and none of the interactions were statistically significant. Age is continuous covariate. Hence, no need to create dummy variables for such covariates and it was included as AGE_i. The final multifactor Cox PH model based on Table: **4.4** is then given by

$$h_i(t) = h_0(t)\exp\begin{pmatrix}1.008WHOCS_{2i} + 1.332WHOCS_{3i} + 2.684WHOCS_{4i} - 0.66CD4_{2i}\\ -10.678CD4_{3i} + 0.302FS_{2i} + 0.221FS_{3i} + 0.76TB_{2i} + 1.51OI_{2i}\\ +0.964RB_{2i} + 0.01AGE_i\end{pmatrix}$$

Once a model has been fitted, there are a number of aspects of the fit of model that need to be studied since model checking is essential part of modeling. Hence, the adequacy of

the fitted model, including the PH assumption and the goodness of fit, was assessed by different diagnostic procedures (See Model Diagnostics for Cox PH Model in section 3.2.4.1). After the values of continuous covariate Age was grouped to give categorical variables, all covariates have been checked for PH assumption by using plot of $\log(-\log(Survival))$ against survival time (Section 3.2.4.2). This plot was roughly parallel for separate groups of patients defined by categories of most covariate. However, there was a little reason to doubt the proportional hazards assumption since there were some covariates which seems to have levels associated with plots that are not in line with others (See Figure7. 2 in the Appendix).

Although such impression results from relatively high cumulative hazard estimates at the longest survival times experienced by patients in that group, these plots take no account of the values of other variables and it could be that the survival times of the individuals in such groups have been affected by the values of other variables. Therefore, other approaches have been also used to certain conclusions. Interactions of the predictors and survival time were included in the model as Time-dependent covariates. There were no evidences that the PH assumption was violated for any of the predictors and it could be noted that covariates became insignificant when the interactions were comprised in the model (Table: 4.5). Likewise, Schoenfeld residuals were used to check the PH assumption separately for each covariate. The p-values for testing whether the correlation between Schoenfeld residuals for each covariate and ranked survival time were all greater than 0.05 except Age for which P-Value=0.0442, which suggested that the PH assumption was plausibly satisfied for all covariates but Age (Appendix, Table7. 2), although the interaction of Age with time was insignificant (Table: 4.5). Consequently, the PH assumption for Age needs to be further assessed. Since it is a continuous covariate, Scaled Schoenfeld residuals for age was plotted against rank of survival time and it suggested that the PH assumption was not violated that much (Appendix, Figure7. 3).

Table: 4.5: Statistical Test for PH Assumptions by Adding Time Varying covariates in to Cox PH Model

Effects of Covariates	_t	Haz. Ratio	P>z	[95% Conf.	Interval]
Main	WHO Clinical Stage				
	II	11.250	0.130	0.684	74.161
	III	55.589	0.054	0.688	737.983
	IV	779.129	0.051	0.362	1352.506
	CD4Percent				
	16-28%	1.989	0.558	0.200	19.792
	Above28%	0.000	1.000	0.000	.
	TB Status	1.906	0.596	0.175	20.714
	Functional Status				
	Ambulatory	0.710	0.471	0.280	1.801
	Bedridden	0.341	0.258	0.053	2.202
	Opportunistic Infections	9.1e+303	.	.	.
	Risk Behavior	117.062	0.116	0.308	44460.756
	Age	0.946	0.103	0.886	1.011
Interaction of covariates with time t	Age	1.001	0.052	1.000	1.001
	Functional Status	1.007	0.164	0.997	1.017
	WHO Clinical Stage	0.986	0.060	0.975	0.998
	CD4Percent	0.985	0.252	0.961	1.011
	TB Status	1.001	0.932	0.975	1.027
	Opportunistic Infections	0.003	1.000	0.000	.
	Risk Behavior	0.961	0.190	0.904	1.020

After checking the PH assumption, Goodness of fit of Cox PH model to the dataset was assessed by residual plots. A plot of the Cox-Snell residuals against the cumulative hazard of Cox-Snell residuals also known as cumulative hazard plot of Cox-Snell residuals was one of the methods used in such diagnostics by residual plots. Hence, the plotted points in (Figure: 4.2) are fairly close to a straight line through the origin, which has approximately unit slope. i.e., the plots seem to have no symmetric departure from a straight line. This suggested that the model fitted to the dataset was seems to be satisfactory. However, a model that fits the dataset needs to be searched since the plots are not perfectly straight line suggesting that further aspects of the model has to be assessed.

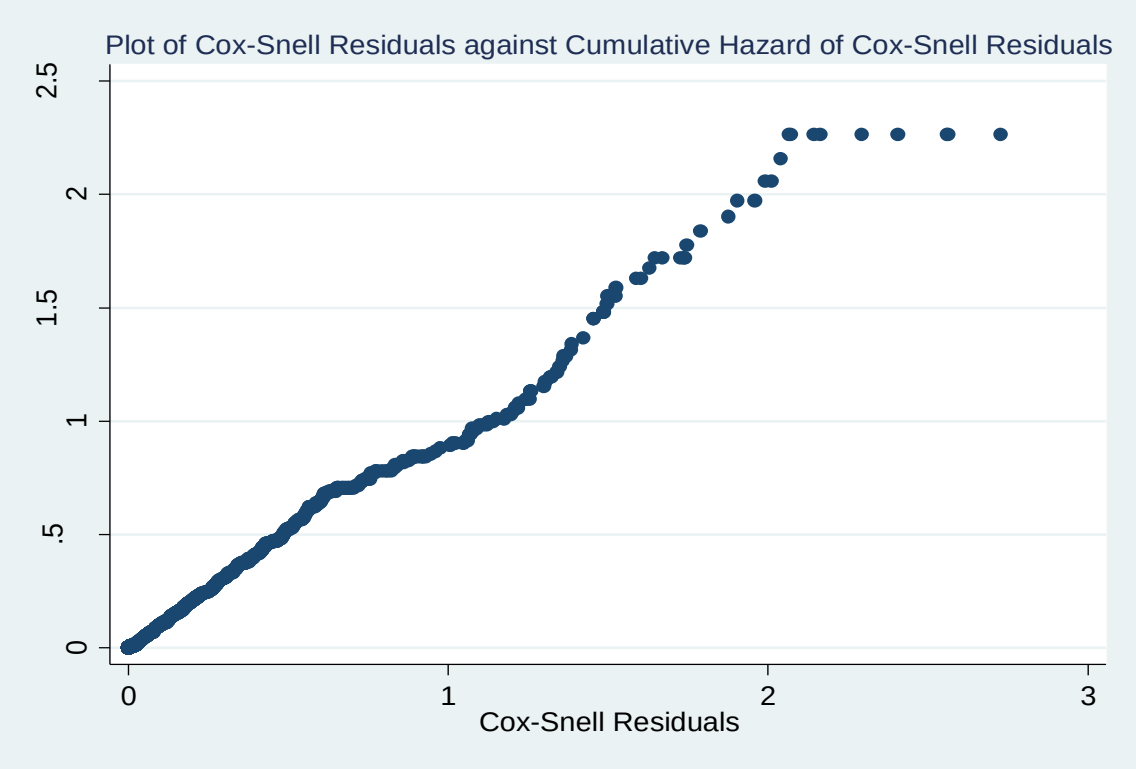


Figure: 4.2: A Cumulative Hazard Plot of Cox-Snell Residuals

Although plot in Figure: 4.2 could help to infer that the Cox PH model fitted to the entire dataset was reasonable, further inspections from different aspects needed to be made in order to increase firmness of the conclusion. As a result, the plot of deviance residuals against the risk scores was used to have information about whether there were patients that might be expected to survive for a short or long time than predicted by the fitted model. Figure: 4.3 shows that the distribution of Risk Score plot of Deviance residuals seems to be approximately symmetrical about zero. This indicates that there are no outlying observations. However, some individuals appear to have smaller value of deviance residual associated with larger risk score (i.e., some individuals seem to survive longer than expected).

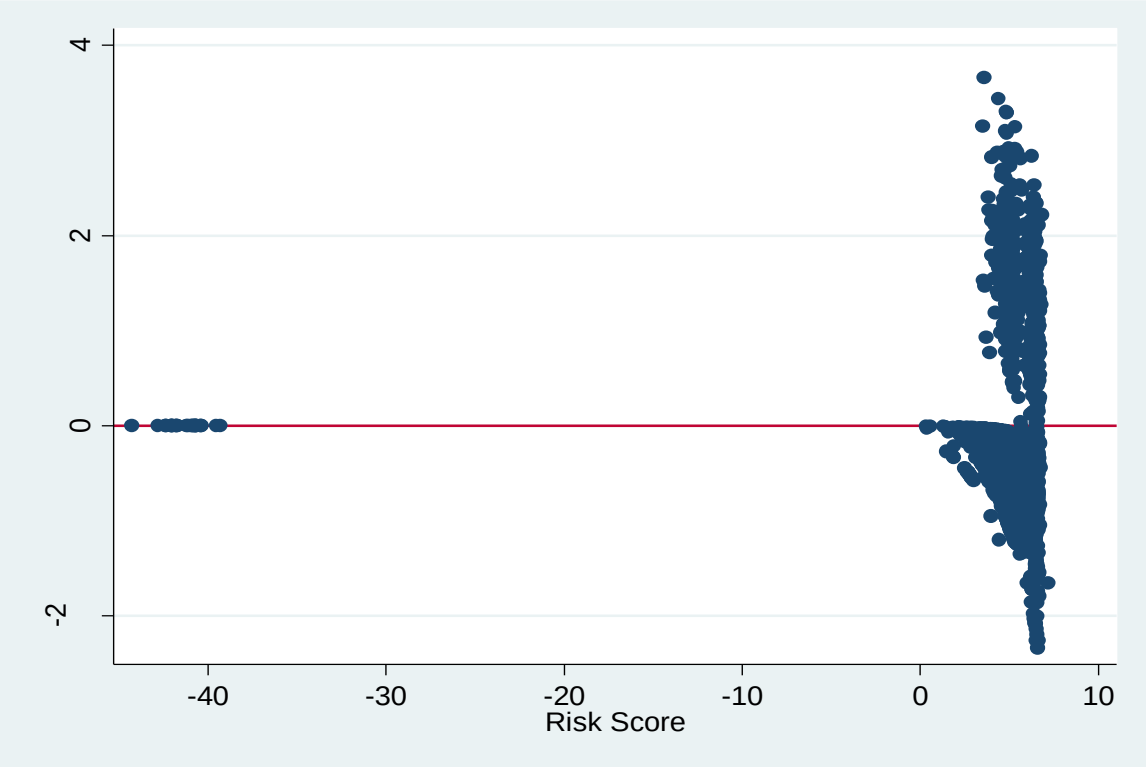


Figure: 4.3: Plot of the Deviance Residuals against the Values of Risk Score

Hence, Delta-Beta (DFBETA) statistics was used to measure the estimated change in coefficients as results of removal of a case since the adequacy of the fitted Cox PH Model needed some reassurance. Index plots of DFBETA statistics for all indicator variables in the Cox PH model were used to detect expected influential observations (Appendix, Figure7. 1). Eventually, on the basis of their standard errors, it has been noted that Coefficients did not change too much when the observations corresponding to the largest Delta-Beta statistics were removed. i.e., magnitudes of all Delta-Beta statistics were less than the corresponding standard error of coefficients (See Table: **4.6**). As a result, such observations were not excluded from the dataset and final conclusion was reached to be there were no influential observations in the dataset.

Table: 4.6: Detecting Expected Outliers by Magnitudes of DFBETA Statistics

Indicators	Expected outlier Index	$\lvert\text{DFBETA}\rvert$	$se(\beta)$
Stage I	Reference	0.000000	
Stage II	2929	0.4992054	0.7373557

Stage III	2929	0.4981006	0.7108191
Stage IV	2929	0.4982148	0.7117566
CD4%: 12-15%	Reference	0.000000	
CD4%: 16-28%	1605	0.039314	0.1719676
CD4%: Above 28%	1375	0.2228134	2.09e+08
FS: Working	Reference	0.000000	
FS: Ambulatory	2120	0.0113839	0.0989787
FS: Bedridden	1380	0.0338106	0.1428324
TB: Negative	Reference	0.000000	
TB: Positive	2684	0.0368193	0.1955617
RB: No	Reference	0.000000	
RB: Yes	108	0.1685705	0.4205107
OIs: No	Reference	0.000000	
OIs: Yes	1827	0.4050735	.7258408
Age	219	0.001903	0.0046558

Furthermore, the continuous covariate Age was checked whether the correct functional form was adopted in the model in order to assess linearity assumption. The Martingale Residual obtained for the fitted Cox PH Model excluding Age was Plotted against the values of Age along with a LOWESS smooth curve (Figure: 4.4). There was no definite pattern in the scatter plots but the smoothed curve deviates from a horizontal line. This indicated that there is a need other than a linear term in age. Thus, the effect of age wasn’t linear in the Cox PH Model fitted to the dataset.

Figure: 4.4: Plot of Martingale Residual against the Values of Age Along with a LOWESS Curve

4.4. AFT Models

The accelerated failure time (AFT) model is another alternative of the Cox PH model when assumptions in the Cox PH are violated. In this particular study, the dataset was fitted using Exponential, Weibull, Log-Logistic, Log-Normal and Generalized Gamma AFT models. In each case, Distribution based stepwise variable selection procedure has been used to identify a set of potential risk factors among covariates that have been significant by the Non-Parametric analysis which results in a good fit of a particular type of AFT Model. As a result, models have been built to possibly minimize loss of information due to exclusion of significant covariates. So that comparison of models on the basis of the dataset will be unbiased. Both independent and multiple covariate analysis were implemented. In the independent analysis, similar conclusions could be made about the effect of individual covariates as it was in the Cox PH model. However, fitted AFT Models seem to provide more information as compared to the fitted Cox PH model since they have additional covariates except Exponential AFT Model. This might

be an indication that AFT Models could be able to identify more factors than Cox PH Model, even if inference is better to made based on statistical criteria (See Table: 4.7).

Although intensive differences have not been examined in estimates of all models and number of covariates were the same in AFT models, possible combination of covariates were different to fit Exponential Regression Model in AFT form. By Stepwise procedures, the continuous covariate Weight was added to all AFT Models over covariates that had been already fitted to Cox PH Model. However, Opportunistic Infections was excluded from Exponential AFT Model instead, Gender has been included to minimize loss of information. As it was in Cox PH Model, Educational Status was not included in any of AFT Models. No interactions were statistically significant in multiple-covariates AFT models. The results from the different AFT models applied to ART dataset are presented in Table: 4.7 and spaces corresponding to covariates which were significant in independent analysis but not selected by stepwise procedure to fit a particular type of AFT Model are shaded.

As a first step to examine Whether an AFT model confer a good fit of the dataset, the Q-Q plot was used for covariates having two categories. The resulting plots approximates well to a straight line from the origin indicating that the AFT model may stipulate an appropriate Model, although there were some plots seem to be far from straight line (Appendix, Figure7. 5).

Table: 4.7 : Multiple-Covariates Analysis: AFT Models

Covariates	Exponential				Weibull				Log-Logistic				Log-Normal				Generalized Gamma			
	β	p	95%CI		β	p	95%CI		β	p	95%CI		β	p	95%CI		β	p	95%CI	
const	10.69	0.00	9.04	12.3	5.37	0.00	5.13	5.62	5.32	0.00	5.09	5.55	5.33	0.00	5.11	5.54	5.35	0.00	5.12	5.58
Age	-0.01	0.02	-0.02	-0.001	-.001	0.02	-.003	-.0002	-.001	0.03	-.003	-.0001	-.001	.04	-.003	-.0001	-.001	0.02	-.003	-.0002
Gender	-0.20	0.04	-0.40	-0.01																
Functional Status (Reference: Working)																				
Ambulatory	-0.32	0.00	-0.52	-0.12	-0.03	0.01	-0.06	-0.01	-0.04	0.01	-0.06	-0.01	-0.04	0.01	-0.07	-0.01	-0.03	0.01	-0.06	-0.01
Bedridden	-0.32	0.03	-0.61	-0.04	-0.02	0.31	-0.05	0.02	-0.02	0.28	-0.06	0.02	-0.03	0.20	-0.07	0.01	-0.02	0.25	-0.06	0.02
Weight	0.01	0.02	0.00	0.03	.001	.04	.0001	.003	.001	0.04	.0001	.003	.002	.03	.0001	.003	.001	0.04	.00008	.003
WHO Clinical Stage (Reference: I)																				
II	-1.12	0.13	-2.57	0.33	-0.12	0.21	-0.30	0.07	-0.10	0.23	-0.26	0.06	-0.07	0.32	-0.19	0.06	-0.09	0.24	-0.25	0.06
III	-1.71	0.02	-3.10	-0.31	-0.16	0.08	-0.33	0.02	-0.13	0.09	-0.29	0.02	-0.09	0.16	-0.21	0.03	-0.13	0.10	-0.28	0.02
IV	-2.84	0.00	-4.24	-1.44	-0.32	0.00	-0.50	-0.15	-0.32	0.00	-0.47	-0.16	-0.28	0.00	-0.41	-0.16	-0.31	0.00	-0.46	-0.16
CD4 Percent (Reference: 12-15%)																				
16-28%	0.75	0.00	0.41	1.09	0.08	0.00	0.04	0.13	0.08	0.00	0.04	0.12	0.08	0.00	0.04	0.12	0.08	0.00	0.04	0.12
Above28%	14.57	0.99	-1548	1577	1.75	0.98	-127	131	1.59	0.98	-148.9	152	1.05	0.98	-95.1	97.2	1.49	0.99	-188.6	191.6
TB Status (Reference: Negative)																				
Positive	-1.11	0.00	-1.48	-0.74	-0.09	0.00	-0.14	-0.04	-0.09	0.00	-0.14	-0.05	-0.09	0.00	-0.14	-0.05	-0.09	0.00	-0.14	-0.05
Education (Reference: Educated)																				
Not Educated																				
OIs (Reference: No)																				
Yes					-0.17	0.01	-0.30	-0.04	-0.19	0.00	-0.31	-0.07	-0.22	0.00	-0.35	-0.08	-0.19	0.00	-0.32	-0.06
RB (Reference: No)																				
Yes	-1.57	0.00	-2.44	-0.70	-0.12	0.02	-0.23	-0.02	-0.12	0.02	-0.23	-0.02	-0.13	0.01	-0.23	-0.03	-0.13	0.02	-0.23	-0.02
/ln_p					2.08	0.00	2.02	2.15												
p					8.03		7.51	8.59												
1/p					0.12		0.12	0.13												
/ln_gam									-2.2		-2.3	-2.2								
gamma									0.11		0.10	0.11								
/ln_sig													-1.59	0.00	-1.65	-1.53	-1.84	0.00	-1.98	-1.71
/kappa																	0.55	0.00	0.31	0.79
sigma													0.20		0.19	0.22	0.16		0.14	0.18

In similar manner with the Cox PH Model, a continuous covariate with a positive coefficient prolongs survival time of HIV/AIDS patient as its value increases. Similarly, level of categorical covariate or an indicator with a positive coefficient is associated with longer survival time of patients as compared to the corresponding reference. On the other hand, covariates/indicators with negative coefficients are associated with accelerated death time as they increase or as they compared to the corresponding references.

Although it would be easy to transform estimated coefficients from PH metric of Exponential and Weibull to their AFT metric by changing the sign of coefficients from Exponential PH and multiplying coefficients from Weibull PH by ancillary parameter $-1/p$ respectively, results of multiple covariate analysis based on stepwise procedure using Exponential and Weibull have been presented in both metrics (See Appendix, Table: 7. 3 for PH metric and Table: 4.7 for AFT metric). Nevertheless, the values of statistical information criteria statistics were the same for AFT and PH form of the model while either of the two distributions was assumed for analysis. Hence, information criteria statistics presented in Table: **4.8** represents not only AFT from of these two but also their PH form as well (See next section).

4.5. Results of Model Comparison

After finding best possible combination of covariates that are assumed to explain the entire dataset with minimum loss of information on the basis of each Model, the performance of models was compared using both $-2l$ and AIC. The $-2l$ was used to compare nested models whereas AIC was used to compare non-nested models. Accordingly, Age, Functional Status, Weight, WHO Clinical Stage, CD4 Percent, TB Status, Opportunistic Infections and Risk Behavior yield the minimum possible loss of information for all parametric models except Exponential AFT Model in which Gender was included instead of Opportunistic Infections (See Table: **4.7**).

Intending to find out the best model for the dataset, Statistical Information Criteria Statistics (. $-2l$ and AIC) were computed for all Models considered in this study (Table: **4.8**). According to the statistics, it could be noted that loss in information reduces as a covariate added to a particular type of model in a stepwise manner. However, the amount

by which the loss reduces does depend on the importance of the covariate being added to the model. The results in Table: **4.8** are therefore presented to make within and between comparison of PH and AFT Models for nested as well as non-nested. Despite, all possible combination of covariates were not fitted. Instead, combination of covariates which thought to be best on the basis of model type specific stepwise procedures since fitting model would add no value and is not statistically meaningful but time consuming.

As a result, Information Criteria Statistics were handled for Cox and Gompertz (as PH Model); Exponential and Weibull (as PH and AFT Model); Log-Logistic, Log-Normal and Generalized Gamma (as AFT Model), starting from null to final models (Table: **4.8**).

Table: 4.8: Statistical Information Criteria for Comparison of Models by Stepwise Selected Covariates for Possible Reduction of Loss of Information

Stepwise Covariate Selection	As PH Models Only				As Both PH and AFT Models				As AFT Models Only					
	Cox		Gompertz		Exponential		Weibull		Log-Logistic		Log-Normal		Generalized Gamma	
	$-2l$	AIC	$-2l$	AIC	$-2l$	AIC	$-2l$	AIC	$-2l$	AIC	$-2l$	AIC	$-2l$	AIC
Null	7217.2	7217.2	1282.8	1286.8	2845.1	2847.1	1228.2	1232.2	1212.0	1215.6	1209.9	1213.9	1207.5	1213.5
WHO Clinical Stage	6931.6	6937.6	975.8	985.8	2562.4	2570.4	930.2	940.2	915.0	925.0	925.2	935.2	911.7	923.7
CD4 Percent	6901.8	6909.8	943.9	957.9	2520.8	2532.8	898.5	912.5	884.7	898.7	896.3	910.3	881.2	897.2
TB Status	6872.8	6882.8	896.2	912.2	2467.3	2481.3	860.4	876.4	855.2	871.2	869.7	885.7	849.3	867.3
Age	6865.3	6877.3	889.9	907.9	2450.1	2466.1	853.7	871.7	849.7	867.7	864.3	882.3	843.1	863.1
Functional Status	6857.2	6875.2	881.4	903.4	2434.4	2454.4	845.9	867.9	840.2	862.2	854.1	876.1	834.3	858.3
Opportunistic Infections	6850.2	6868.2	863.5	887.5			834.4	858.4	827.5	851.5	841.9	865.9	822.3	848.3
Risk Behavior	6843.1	6863.1	857.2	883.2	2421.3	2443.3	828.5	854.5	822.0	848.0	836.2	862.2	816.6	844.6
Weight			852.6	880.6	2418.0	2442.0	824.3	852.3	817.9	845.9	831.7	859.7	812.3	842.3
Gender					2413.6	2439.6								

Therefore, based on the results in Table: **4.8**, all models were compared using Statistical Information Criteria (AIC and $-2l$) for stepwise selected covariates. Since we had a little concern about the fit of Cox PH Model, all parametric models appeared to fit better than Cox PH. Being nested with in Gamma Model, the Exponential, Weibull, and Log-Normal models were compared using the statistic $-2l$. Hence, starting from null to adding the third covariate step wise, the Log-Normal model appeared with the minimum statistic and seemed to fit better. But the Weibull Model fitted the dataset better since it is associated with minimum statistic at the final step. However, the statistic $-2l$ is not valid for comparing models that are not nested. In such case, the statistic AIC was used to compare the models. Although, the Log-Logistic Model is not nested in Gamma, it was better than Weibull. The Gompertz Model provided better fit as compared to Cox PH and Exponential Models.

As a final point, Generalized Gamma Model appeared to be an appropriate AFT model according to AIC compared with other models, although it is only slightly better than Log-Logistic and Weibull (as PH and AFT) models. It has also been noted that the Cox PH and Exponential (as PH and AFT) models were poorer fits according to Information Criteria Statistics suggesting that assuming exponential distribution for the data set would be highly misleading. Thus, this provided more evidence that could strengthen the doubt on PH assumption since the 1st (Gamma) and 2nd (Log-Logistic) best Models were AFT form rather than PH.

After identifying the best model in identifying factors affecting the survival of HIV/AIDS patients under ART follow-up, the goodness of fit of the model was checked using residual plots. The cumulative hazard plot of the Cox-Snell residuals in Generalized Gamma model is presented in Figure: 4.5. The plotted points lie on a line that has a unit slope and zero intercept. So there was no reason to doubt the suitability of this fitted Generalized Gamma model. Eventually, the conclusion appeared to be the Generalized Gamma model was the best and perfect fitting AFT model based on AIC criteria and residuals plot.

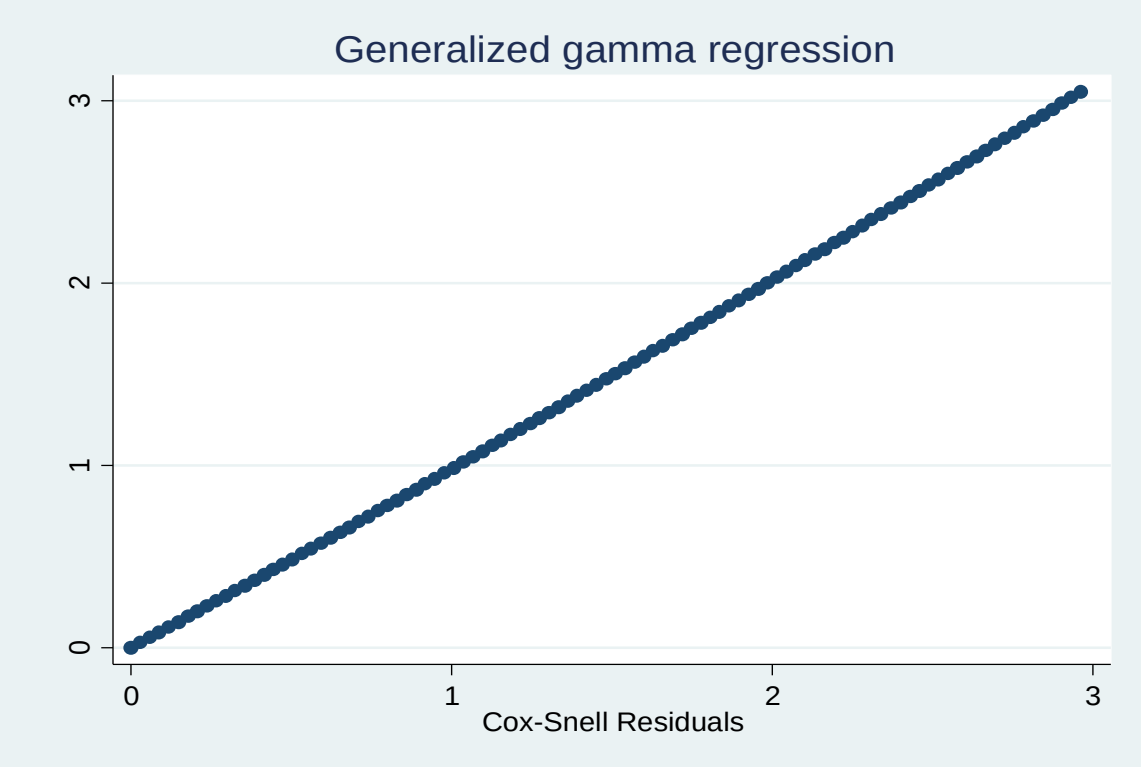


Figure: 4.5: Cumulative Hazard Plot of the Cox-Snell Residual for Generalized Gamma AFT Model

The best fitting survival model for the dataset is Generalized Gamma AFT Model. Henceforth, the effect of covariates will be interpreted in terms of Time Ratios based on the best fitted model in order to understand it in easier way. In STATA, the Reporting option, tr specifies that exponentiated coefficients, which are interpreted as time ratios and it is appropriate only for the Log-Logistic, Log-Normal, and Gamma models, or for the Exponential and Weibull models when fit in the accelerated failure-time metric (tr may be specified at estimation or upon replay). Therefore, Time Ratios associated with selected significant covariates that represented the entire dataset on the basis of Generalized Gamma AFT Model are presented in Table: 4.9 and the corresponding coefficients are presented in Table: **4.7** together with other models considered in the study.

Table: 4.9: Time Ratios on the Basis of Generalized Gamma AFT Model

Covariates	Tm. Ratio	Std. Err.	Z	P>z	[95% Conf. Interval]	
Age	.999	.001	-2.28	0.02	.997	.9998
Functional Status (Reference: Working with time ratio=1)						
Ambulatory	.966	.013	-2.61	0.01	.941	.9914
Bedridden	.977	.020	-1.16	0.25	.940	1.016
Weight	1.0014	.001	2.07	0.04	1.0001	1.003
WHO Clinical Stage (Reference: I with time ratio=1)						
II	.911	.072	-1.18	0.24	.780	1.064
III	.882	.067	-1.65	0.10	.759	1.024
IV	.734	.056	-4.02	0.00	.632	.8534
CD4 Percent(Reference: 12-15% with time ratio=1)						
16-28%	1.084	.023	3.81	0.00	1.040	1.1297
Above28%	4.448	431.4	0.02	0.99	1.25e-82	1.58e+83
TB Status(Reference: Negative with time ratio=1)						
Positive	.912	.022	-3.86	0.00	.870	.9555
Opportunistic Infections(Reference: No with time ratio=1)						
Yes	.826	.055	-2.85	0.00	.725	.9420
Risk Behavior(Reference: No with time ratio=1)						
Yes	.882	.046	-2.39	0.02	.796	.9776
_cons	211.4	24.9	45.40	0.00	167.8	266.4
/ln_sig	-1.844	.0705	-26.15	0.00	-1.982	-1.705
/kappa	.552	.122	4.54	0.00	.314	.7907
Sigma	.158	.0112			.138	.1817

Under the Generalized Gamma AFT model, the estimated acceleration factor for a year increase in age considering the effect of other covariates in the model constant was 0.999 with 95%CI: [0.997, 0.9998], which indicates that older patients were more likely to die earlier than patients who were relatively younger. In similar fashion, it was 1.0014; 95%CI: [1.0001, 1.003], for a Kg increase in weight providing an evidence that patients with relatively larger weights were more likely to survive as compared to patients with relatively smaller Weights (Table: 4.9). The corresponding coefficients can also be interpreted as additive effects on the Natural Log of survival time due to a unit increase in that particular continuous covariate in question holding the effects of other covariates constant (Table: **4.7**).

Analogously, the estimated acceleration factors for Ambulatory and Bedridden patients were 0.966; 95%CI: [0.941, 0.9914], and 0.977; 95%CI: [0.940, 1.016], respectively as compared to working patients suggesting that the survival time of working patients was

longer than patients with functional status Ambulatory and Bedridden. However, the indicator associated with Bedridden was insignificant (Table: 4.9). In terms of coefficients, the Natural Log of survival times for Ambulatory and bedridden patients decrease by 0.03 and 0.02 respectively as compared to the Natural Log of survival times for working patients[2]. Furthermore, WHO Clinical Stage was also one of the significant covariates with acceleration factor 0.911; 95%CI: [0.780, 1.064], 0.882; 95%CI: [0.759, 1.024] and 0.734; 95%CI: [0.632, 8.534] for Stage II, III and IV patients respectively. This indicates that although Stage II and III were not significant, patients with higher WHO Clinical Stages were less likely to survive longer as compared to that of with Stage I. It has also been noted that, patients with CD4 percent between 16 and 28 significantly survived longer than patients having CD4 percent between 12 and 15 with estimated acceleration factor 1.084. The estimated median survival time for patients having CD4 percent above 28 was 4.448 times that of patients having CD4 percent between 12 and 15. However, it was not significant due to absence of recorded death event in that group (Table: 4.9).

Not only these, but also being TB positive was one of significant factors that shortened the life span of HIV patients. According to the results in Table: 4.7, estimated average survival time of TB positive patients under Natural logarithm was less by 0.09 as compared to TB negative patients. In other words, being TB positive was associated with accelerated death time with estimated acceleration factor 0.912 indicating that TB Positive patients have approximately 9% less average survival time than TB Negative patients. Likewise, the estimated acceleration factor for patients having any kind of risk behavior was 0.882 and it was 0.826 for patients with any opportunistic infection (Table: 4.9). This was evidence that having risk behavior or opportunistic infection lets patients die earlier than patients who do not have any risky behavior or opportunistic infections. The decreases in the natural logarithm of survival time were 0.13 (patients with risky behavior) and 0.19 (patients having OIs) as compared to patients with no risky behavior and patients do not have OIs respectively (Table: **4.7**).

[2] *Note that, the effect of a particular concomitant factor on Patients Survival from a model fitted to more than one covariate is interpreted by considering the effects of the rest covariates in the model as constant.*

Now, Model based predictions need to be derived and from equation [3.55], the time ratio or median survival time for a particular patient (i^{th} individual) with authentic baseline values of selected covariates relative to the median survival time computed from baseline survivor function which takes zero (reference) values of all covariates under the Generalized Gamma model is given by

$$TR_{iGG}(50) = \exp\begin{pmatrix} -0.001AGE_i - 0.03FS_{1i} - 0.02FS_{2i} + 0.01WEIGHT_i - 0.09WHOCS_{1i} \\ -0.13WHOCS_{2i} - 0.31WHOCS_{3i} + 0.08CD4_{1i} + 1.49CD4_{2i} - 0.09TBS_{1i} \\ -0.19OIs_{1i} - 0.13RB_{1i} \end{pmatrix}$$

Hence, the relative median survival time for an individual patient can be predicted with the help of this model.

Note that, the shape parameter k (*kappa*) tested whether it is significantly different zero and one to check the appropriateness of Log-Normal and Weibull AFT Models respectively. The corresponding Wald tests provided that the shape parameter is significantly different from zero and one as well. From Table: **4.7**, the 95% CI for k was [0.31, 0.79] with $p - value = 0.000$ indicating that the Log-Normal AFT model is not appropriate. Similarly, after fitting the full Generalized Gamma AFT Model, the command (test [kappa]_cons=1) in STATA resulted in Chi-Squared value of 13.55 with 1 degree of freedom and $p - value = 0.0002$ suggesting that the Weibull AFT Model was not appropriate model for the dataset. Hence, the Generalized Gamma AFT model that allows variety of shapes of the hazard function through time was the best fit to the dataset in every aspect of model comparison since the change in hazard was not monotonic over time.

4.6. Discussion

The two often-used models for adjusting survivor functions for the effects of covariates are the accelerated failure-time (AFT) model and the multiplicative or proportional hazards (PH) model. However, Cox PH Model probably seems to be interest of most medical researchers than the AFT metric in which survival time is assumed to follow a particular distribution. In using the Cox PH Model, proportional hazards assumption needs to be satisfied to make trustworthy inference from a particular dataset. But the PH assumption may be violated for some reason based on the nature of the dataset. Although adding time dependent covariates and taking the covariate associated with violence of PH assumption as stratification id variable (often for categorical covariates) are solutions, in such case, AFT Model is more informative and the best alternative to use since it can provide precise estimates if distributional assumption satisfied. Moreover, interpretation of covariates effect from AFT Model is directly associated with average survival times unlike Cox PH Model in which estimates are interpreted as multiplicative effect on hazards.

The parametric distributions Exponential, Weibull, Gompertz, Log-Logistic, Log-Normal and Generalized Gamma are most commonly encountered in survival analysis and are currently supported by software like STATA. Exponential and Weibull distributions can be used in both metric. However, the Gompertz distribution is applicable only in PH metric whereas the rest in AFT metric only. As a result, some studies were found to be conducted on comparison of the two metric. However, inconsistencies arisen in decisions made on appraisal from different datasets. This is probably due to the nature of the event that generates the time. In spite of this, model wise variable selection procedures need to be performed that might be responsible for different set of covariates. But most studies conducted regardless of model wise selection procedures (See [34], [45]). In such case, statistical information criteria for a particular model containing insignificant covariate (s) will be inappropriately large if there are covariates in the dataset that can be significantly fitted by the model. Hence, the model on the basis of which the selection procedure is made will be favored to represent the data set unbecomingly and misleading conclusion will be made on appraisal. In fact, it is one of the problems that revived this study.

Although there were studies with a final conclusion that the Cox PH Model provided a better fit than AFT Models [11], according to conclusions made by most of the researchers, it is customary that AFT models were best based on statistical criteria (See section 2.2.2). In studies conducted in PH metric, The Gompertz distribution is relatively less used. In contrast, Exponential and Weibull models were found to be more used and likely to fit a particular dataset than Cox PH Model (See [41] and [57]). Log-Logistic AFT Model was frequently chosen over the other when comparison between the two metric is considered. However, Log-Normal and Generalized Gamma AFT Models were also preferred by some researchers on the basis of information criteria.

In this particular study, the Generalized Gamma AFT Model was the best model in determining maximum possible number of factors affecting the survival of HIV/AIDS patients under ART follow-up with relatively minimum loss of information. Fittingly, patients' WHO Clinical Stage, CD4 Percent, TB Status, Functional Status, Age, Risk Behavior, Weight and Opportunistic Infections were identified to have significant effect on survival time of HIV patients. The Logistic frame work and Cox PH models were often used methods in determining such factors (See section 2.2.1). Yet, survival analysis is more appropriate for such inquiry preferably, the AFT metric. Since availability of software packages made it easy to choose the most appropriate model from obtainable survival models, we don't need prior assumption regarding specific type of distribution for event times instead, analyzing the data by using a model with significantly minimum information criteria statistic. Therefore, inference could be made in easily understandable and sensible way besides get hold of more precise and reliable estimates of effects. Consequently, there will probably high chance of minimizing contradictions on conclusions made from different researchers on a particular area.

As specific aim of the study, the effect Age, Gender and Level of Education was tried to be assessed to get better confidence to wards inference on their effect. Although they were all significant independently, Age was the only significant factor in multiple-covariates analysis. Hence, in the presence of other significant covariates considered in the study, the conviction that Gender and Level of Education don't significantly affect the survival of HIV patients but Age is strengthened. This was true for all parametric Models

the explored the same set of covariates except Exponential regression Models that have relatively high information criteria statistics as compared to the other parametric models.

5. Conclusions and Recommendations

5.1. Conclusions

Despite the challenges, this study was conducted to investigate the effects of socio-economic, demographic and health factors on survival of HIV/AIDS patients. According to information criteria statistics (AIC and $-2l$), the Generalized Gamma AFT model best described the dataset on a large number of patients from University of Gondar teaching Hospital. Using stepwise selection procedures for Generalized Gamma AFT model, WHO Clinical Stage, CD4 Percent, TB Status, Functional Status, Age, Risk Behavior, Weight and Opportunistic Infections were fitted as best possible combination of covariates, although Gender and Level of Education were ascertained to be potential risk factors by Non-Parametric as well as independent survival analysis. On the other hand, Marital Status Occupational Status and Cotrimoxazol were insignificant at 5% level of significant by Non-Parametric analysis. Hence, the stepwise procedures were performed with chance of 10% exclusion and 5% inclusion from/to a model for only 10 covariates known to be significant by Non-Parametric or independent analysis. According to the multiple-covariates analysis of the best fitted model, the study revealed that advanced WHO clinical stages (III and IV), lower CD4 percent (12-15%), TB co-infection, being bedridden or ambulatory functional status, being relatively old in age, having relatively lower weight, the presence of Opportunistic Infections and Risky Behaviors were strongly related to relatively minimum average survival time or accelerated death time.

Furthermore, Log-Logistic, Weibull, Log-Normal and Gompertz regression models were the 2^{nd}, 3^{rd}, $4^{th}$and 5^{th} best model by statistical model comparison criteria whereas the Exponential Model was ranked as the 6^{th} model in describing the data, although the Cox PH model provided the poorest fit due to some concern about its fit probably as a result of violence in the PH assumption.

5.2. Recommendations

- Researchers who are interested in determining factors affecting survival of HIV/AIDS patients can explore the effects of some other factors such as Gender and Level of Education if they fit models based on purposely selected significant covariates.
- Health workers should be cautious when a patient is in advanced clinical stages, old in age, relatively lower in weight, bedridden, opportunistically infected, TB positive, hazardous in behavior and has lower CD4 percent during ART initiation.
- Concerned government authority/Health policy makers should arrange Nation wise trainings that can provide health workers and data clerks with uniform handling of information related to patients and law-abiding delivery of secured quality data for stakeholders at all level.

Bibliography

[1]. **Aalen O, Borgan O, Gjessing H.** *Survival and Event History Analysis.* New York : Springer-Verlag, 2008.

[2]. **Alinda M. Bosch, Frans J. Willekens, Abdullah H. Baqui, Jeroen K. S. Van GinnekenInge Hutter.** *Association between Age at Menarche and Early-Life Nutritional Status in Rural Bangladesh* . s.l. : J.Biosoc.Sci, 2008 . 40, 223–237.

[3]. **Altshuler, B.** *Theory for the Measurement of Competing Risks in Animal Experiments.* . s.l. : Math Bioscience, 1970. 6: 1 – 11.

[4]. **Anderson, P. K., Borgan, O., and Gill, R. D.** *Coxís regression model counting process: A large sample study.* . s.l. : Annals of Statistics, 1982. 10,1100-1120..

[5]. **Barlow, W. E., and Prentice, R. L.** *Residuals for relative risk regression.* . s.l. : Biometrika, 1988. 75 , 65-74..

[6]. **Belaynew Wasie, Yigzaw Kebede, Anwar Yibrie.** *Nutritional Status of Adults Living With HIV/AIDS at the University of Gondar Referral Hospital, Northwest Ethiopia* . s.l. : Ethiop. J. Health Biomed Sci, 2010. 3(1).

[7]. **Carvour, Martha Lydia.** *Patterns and Predictors of Survival Following an HIV/AIDS-Related Neurologic Diagnosis Phd (Doctor of Philosophy) Thesis, .* s.l. : University Of Iowa, 2012. Http://Ir. Uiowa.Edu/Etd/2454.

[8]. **(CDC), Centers for Disease Control and Prevention.** *1993 Revised Classification System For HIV Infection And Expanded Surveillance Case Definition For AIDS Among Adolescents And Adults.* s.l. : Morbidity and Mortality Weekly Report (MMWR) , 1992. 41, 17.

[9]. **Story, Dr. David Ho: Person of the Year.** *Chua-Eoan, Howard.* s.l. : Time, 1996.

[10]. **Collett, D.** *Modelling Survival Data in Medical Research (Second Edition).* London : Chapman And Hall/CRC, 2003.

[11]. **Conge C, Tsoikos CP.** *Statistical Modeling Ofbreast Cancer Relapse Time with Different Treatments. Journal of Applied Sciences.* 2010. 10:37-44.

[12]. **Cox, D. R., and Snell, E. J.** *A general definition of residuals with discussion. Journal of the Royal Statistical Society. Series B.* 1968. 30, 248-275.

[13]. *Journal of the Royal Statistical Society.* **Cox, D.R.** Regression Models and Life Tables (With Discussion), 1972, Vol. 34. 187-220.

[14]. *J.Biosoc.Sci.* **Alinda M. Bosch, Frans J. Willekens, Abdullah H. Baqui, Jeroen K. S. Van GinnekenInge Hutter:.** . Association between Age at Menarche and Early-Life Nutritional Status in Rural Bangladesh . s.l. , 2008 , Vol. 40. 223–237.

[15]. **Diez, David M.** *Survival Analysis In R Open intro* . s.l. : Openintro.Org, 2013.

[16]. **David Ross, Bruce Dick, Jane Ferguson.** *[Preventing HIV/AIDS In young People: A Systematic Review of the Evidence From developing Countries: UNAIDS Interagency Task Team on HIV and Young people.* s.l. : WHO Technical Report Series. 938.

[17]. **Davis, Davis.** *Tree-Augmented Cox Proportional Hazards Models.* 2009.

[18]. **Derek N. N., Albert L., Timothy A.** *Performance of Cox Proportional Hazard and Accelerated Failure Time Models in the Analysis of HIV/TB Co-infection Survival Data* . 2014. 4 , 94-102.

[19]. *The Journal of Test Positive Aware Network.* **Diaz-Linares, Mariela, and Enid Vázquez.** 12th Annual HIV Drug Guide, 2008.

[20]. **Wenyuwang, Elisa T. Lee John.** *Statistical Methods For Survival Data Analysis Third Edition* . s.l. : John Wiley & Sons, Inc., 2003.

[21]. **Engel, Jonathan.** *The Epidemic: A Global History of AIDS.* New York : Smithsonian/Collins., 2006.

[22]. **Ethiopia, Federal Democratic Republic Of.** *Country Progress Report On The HIV Response.* . 2014.

[23]. **Office, Federal HIV/AIDS Prevention and Control.** *Federal Ministry of Health Strategic Plan for intensifying multi-sectorial HIV and AIDS Response in Ethiopia* . Addis Ababa, Ethiopia : s.n., 2009 – 2014. II (Spm II).

[24]. **Research., Food and Drug Administration (FDA)/Center for Biologics Evaluation and.** *Testing Your-self For HIV-1, The Virus That Causes AIDS." FDA Web Site. Http://Www.Fda.Gov/Cber/Infosheets/Hiv-Home2.Htm.* 2008.

[25]. *The Journal of Test Positive Aware Network.* **Gallant, Joel.** HIV Drug Guide Introduction, 2008.

[26]. *British Journal Of Medicine & Medical Research.* **Gemeda Bedaso Buta, Ayele Taye Goshu and Hailemichael M. Worku.** Bayesian Joint Modelling Of Disease Progression Marker And Time To Death Event Of HIV/AIDS Patients Under ART Follow-Up, 5(8): 1034-1043, 2.

[27]. *Global Report: UNAIDS Report On The Global AIDS Epidemic.* 2013.

[28]. **Goldman, Bonnie.** *Three Cases of HIV Transmission to Infants through Food Pre-Chewedby HIV-Positive Caregiver.* 2008.

[29]. **Mekonnen, Hailu Tadeg and Negussu.** *Trends in Antiretroviral Drugs Prescribing At Public Health Facilities in Ethiopia: Compliance to Treatment Guidelines. .*

[30]. **(HAPCO), HIV/AIDS Prevention and Control Office.** *Multi-sectorial Plan Of Action For Universal Access To HIV Prevention, Treatment, Care And Support In Ethiopia.* 2007 –2010.

[31]. **Hosmer, D.W. And Lemeshow S.** *Applied Survival Analysis. .* New York : John Wiley and Sons, Inc, 1999.

[32]. **Isidore Sieleunou, Mohamadou Souleymanou, Anne-Marie Scho ¨Nenberger, Joris Menten And Marleen Boelaert.** *Determinants of Survival in AIDS Patients on Antiretroviral Therapy in A Rural Centre In The Far-North Province, Cameroon. .* Volume 14 No 1 Pp 36.

[33]. **Jane Lu, Astrazeneca Pharmaceuticals, Wilmington, DE David Shen,.** *Independent Consultant Survival Analysis Approaches and New Developments Using SAS. Pharmasug .* 2014. - Paper PO02.

[34]. *International Journal of Data Envelopment Anual.* **Jemal Ayalew, Helen Moges, Omprakash Sahu, And Anteneh Worku,.** Identifying Factors Related To The Survival Of AIDS Patients Under The Follow-Up Of Antiretroviral Therapy (ART): The Case Of South Wollo,

[35]. **Qi, Jiezhi.** *Comparison of Proportional Hazards and Accelerated Failure Time Model.* Mar. 2009.

[36]. **Sons, John Wiley &.** *Mathematical Methods In Survival Analysis, Reliability And Quality Of Life.* s.l. : John Wiley & Sons Inc, 2008.

[37]. —. *Statistical Advances in the Biomedical Sciences Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics.* s.l. : John Wiley & Sons, Inc. , 2008.

[38]. **Kahn, Patricia, Ed.** *AIDS Vaccine Handbook: Global Perspectives. : AIDS Vaccine Advisory Coalition.* New York : s.n., 2005.

[39]. **Glass, Kathy S. Stolley and John E.** *HIV/AIDS: Health and Medical Issues Today.* 2009.

[40]. **Kebebew, Ketema.** *Determining Factors That Affect The Survival Rate Of HIV-Infected Patients On Art: The Case Of Armed Forces General Teaching Hospital.* Addis Ababa, Ethiopia. : s.n., JUNE, 2011.

[41]. **Kleinbaum D, Klein M.** *Survival Analysis: A Self- Learning Text,, Springerverlag.* New York : s.n., 2005.

[42]. **Lumbwe Chola, Lars T. Fadnes, Ingunn M. S. Engebretsen, James K. Tumwine, Thorkild Tylleskar, Bjarne Robberstad, and the PROMISE EBF Study Group.** *Infant Feeding Survival and Markov Transition Probabilities Among Children Under Age 6 Months: Practice of Epidemiology.* 2013.

[43]. **Mekonnen, Yared, Rachel Sanders, Senait Tibebu, and Priya Emmart.** *Equity and Access To ART In Ethiopia. Futures Group, Health Policy Initiative, Task Order 1.* Washington, DC : s.n., 2010.

[44]. **Miguel A. Herna´N, Stephen R. Cole , Joseph Margolick , Mardge Cohen And James M. Robins.** *Structural Accelerated Failure Time Models for Survival Analysis In Studies With Time-Varying Treatments.* 2005.

[45]. **Health, Ministry Of.** *Guideline for Implementation of Antiretroviral Therapy in Ethiopia. .* January, 2005.

[46]. **Ayaneh, Muluye Getie.** *Modelling Time-To-Menarche: Cox Proportional Hazard versus Accelerated Failure Time Models.* October, 2011.

[47]. **(NIAID), National Institute Of Allergy and Infectious Diseases.** *Challenges In Designing HIV Vaccines.* 2008a.

[48]. **(NIAID), National Institute Of Allergy and Infectious Diseases.** *Treatment of HIV Infection.* 2008b.

[49]. *Journal of Quality Technology,*. **Nelson, W.** Hazard Plotting For Incomplete Failure Data, 1969. 1: 27 – 52.

[50]. **Nelson, W.** *Theory And Application Of Hazard Plotting For Censored Failure Data.* s.l. : Technometrics, 1972. 14: 945 – 9685.

[51]. **Ibrahim, Nuredin.** *Evaluation of Factors Affecting the Chance of Survival/Death Status among HIV Positive People under the antiretroviral treatment Program: The Case of Adama Hospital.* AUGUST 2007.

[52]. **Gilbert., Peter B.** *Failure Time Analysis Of HIV Vaccine Effects On Viral Load And Antiretroviral Therapy Initiation.* . April 14, 2005.

[53]. *Indian Journal of Science and Technology.* . **Ponnuraja C, Venkatesan P.** Survival Models for Exploring Tuberculosis Clinical Trial Dataan Empirical Comparison, 2010. 3: 755-58..

[54]. *Asian Pacific J Cancer Prev.* **Pourhoseingholi M .A, Hajizadeh E., Moghimi B.Et Al.** Comparing Cox Regression And Parametric Models For Survival Ofpatients With Gastric Carcinoma., 2007. 8, 412-416..

[55]. *Iran J Cancer Prev.* **Pourhoseingholi MA, Pourhoseingholi A, Vahedi M, Et Al.** Alternative for the Cox Regression Model: Using Parametric Models to Analyze the Survival of Cancer Patients, 2011. 4:1- 9..

[56]. *Open Journal Of Statistics,*. **Radhey S. Singh, Dishna P. Totawattage.** The Statistical Analysis of Interval-Censored Failure Time Data with Applications., 2013. 3, 155-166.

[57]. **Richard Williams.** *Alternatives to Logistic Regression (Brief Overview),*. 2015.

[58]. *Journal Of Experimen& Clinical Cancer Research.* **Sayehmiri K, Eshraghian MR, Mohammad K, et al.** Prognostic Factors Of Survival Time After Hematopoietic Stem Cell Transplant Inacute Lymphoblastic Leukemia Patients: Cox Proportional Hazard Versus Accelerated Failure Time Models., 2008. 27:1-9..

[59]. **Schoenfeld, D.** *Chi-Squared Goodness-Of-Fit Tests For the Proportional Hazards Regression Model.* s.l. : Biometrika, 1982. 7: 145-153..

[60]. **Seth C. Kalichman, CT Kluwer, et al.** *Positive Prevention Reducing HIV Transmission Among People Living With HIV/AIDS* . 2005.

[61]. **Shah, Sonia.** *The Body Hunters: Testing New Drugs on the World's Poorest Patients.* . New York : New Press, 2006.

[62]. *International Journal of Science and Research.* . **Shankar Prasad Khanal, V. Sreenivas, Subrat K. Acharya.** Accelerated Failure Time Models: An Application In The Survival Of Acute Liver Failure Patients In India, June, 2014, Vol. 3.

[63]. **Smith, Kendall A.** *"The HIV Vaccine Saga."Medical Immunology.* 2003. 2, 1.

[64]. **Selvin, Steve.** *Survival Analysis For Epidemiologic And Medical Research A Practical Guide.* 2008.

[65]. **Stine, Gerald J.** *AIDS Update 2008. Boston: Mcgraw-Hill.* 2009.

[66]. **WR., Swindell.** *Accelerated Failure Time Models Provide A Useful Statistical Framework For Aging Research. Experimental Gerontology.* 2009 . 44: 190–200..

[67]. **Müller, Tanja R.** *HIV/AIDS, Gender and Rural Livelihoods in Sub-Saharan Africa an Overview and Annotated Bibliography* . 2005. AWLAE SERIES No. 2..

[68]. *Ethiop. J. Health Dev.* . **Tegiste Assefa, Eshetu Wencheko.** Survival Analysis Of Patients Under Chronic HIV-Care And Antiretroviral Treatment At Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia : s.n., 2012. 26(1):22-29.

[69]. **AIDS, The 2011 United Nations Political Declaration on HIV and.** *Global Aids Response Progress Reporting 2014 construction of Core Indicators for Monitoring.* 2014.

[70]. **UNITAID, The Clinton Health Access Initiative and.** *HIV/AIDS in Ethiopia.* August, 2011.

[71]. **Therneau, T. M., Grambsch, P. M., and Fleming, T. R.** *Martingale-based residuals for survival models.* s.l. : Biometrika , 1990. 77 , 147-160.

[72]. **UNAIDS.** *Fact Sheet.* 2014.

[73]. *AIDS By the Numbers.* 2013.

[74]. **Africa, WHO Regional Office for.** *Ethiopia Factsheets of Health Statistics* . 2010.

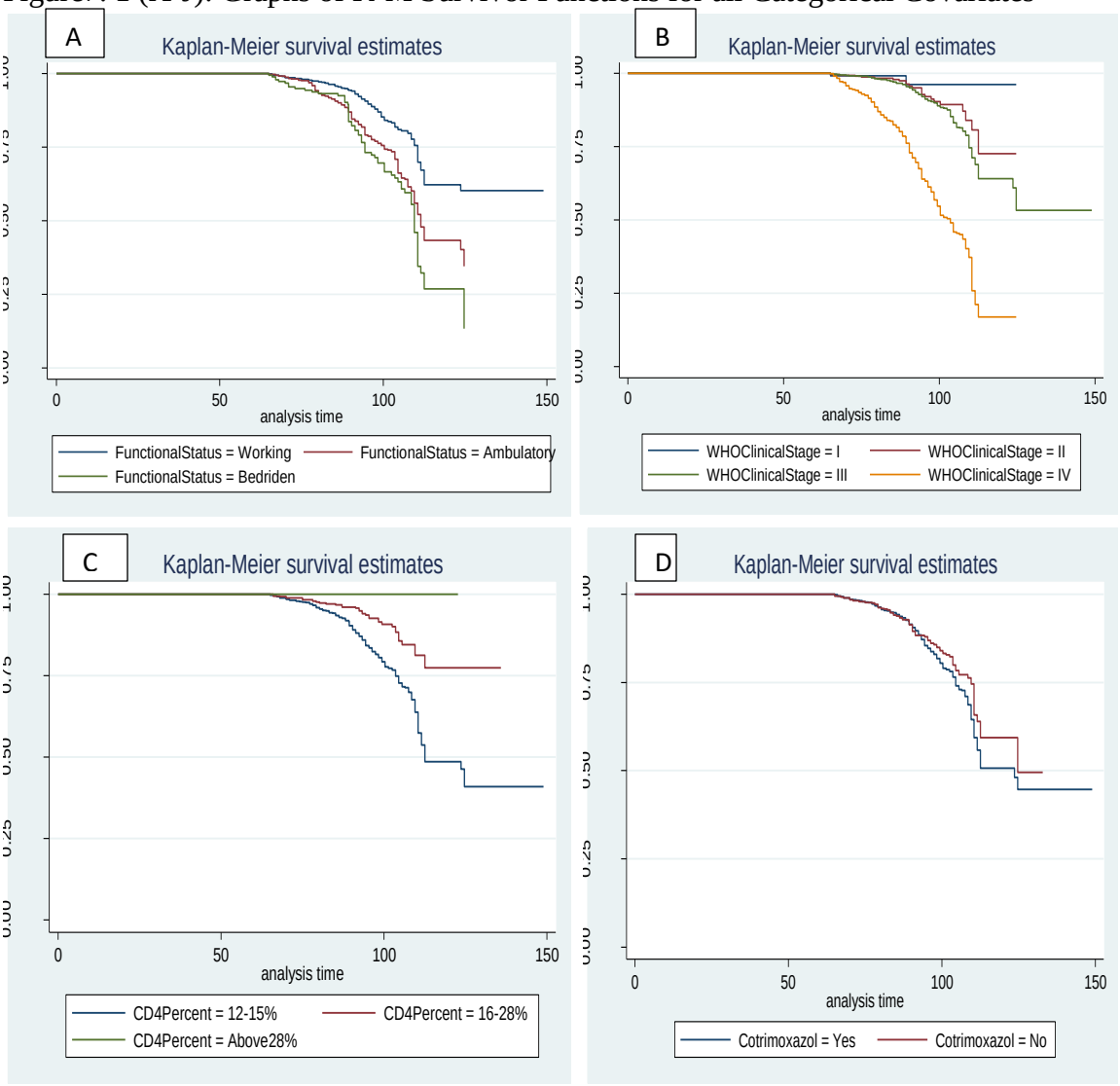
Appendix

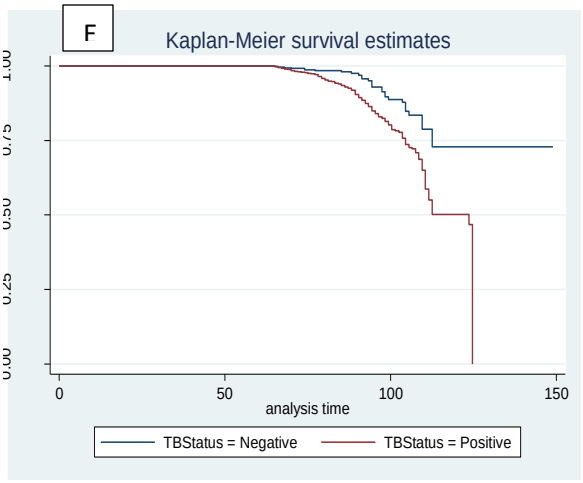
Summary Statistics of Continuous Covariates and Graphs of K-M Survivor Functions for Categorical Covariates

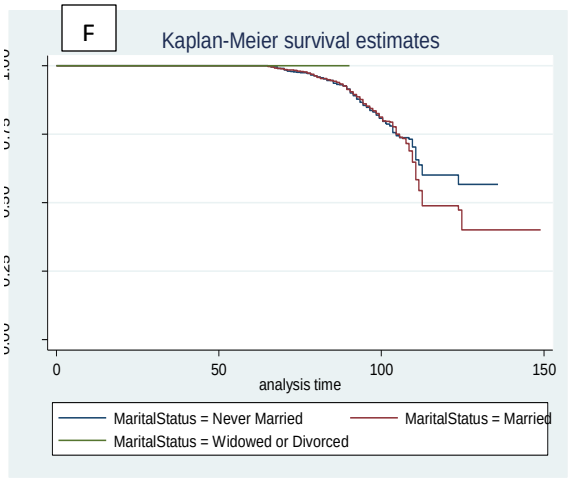
Table7. 1: Summary Statistics for Continuous Covariates in the Dataset

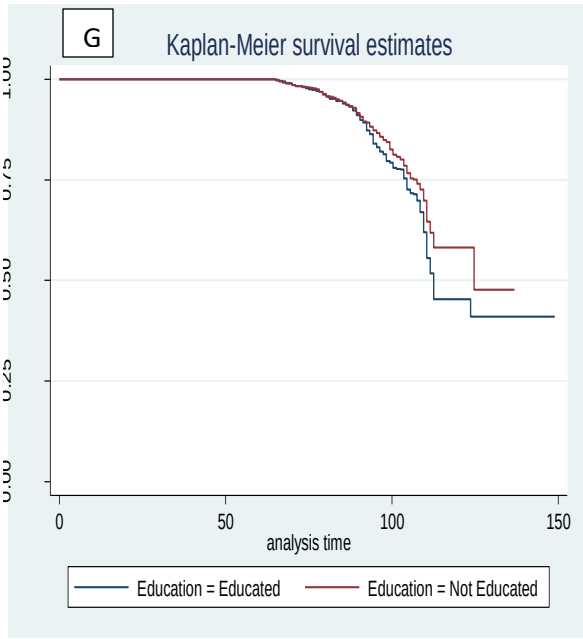
Variable	Observations	Mean	Std. Dev.	Min	Max
Age	3042	32.37881	11.12487	.03	99
Weight	3042	46.03455	11.3793	3	89
Household Size	3042	3.626233	1.560996	1	10

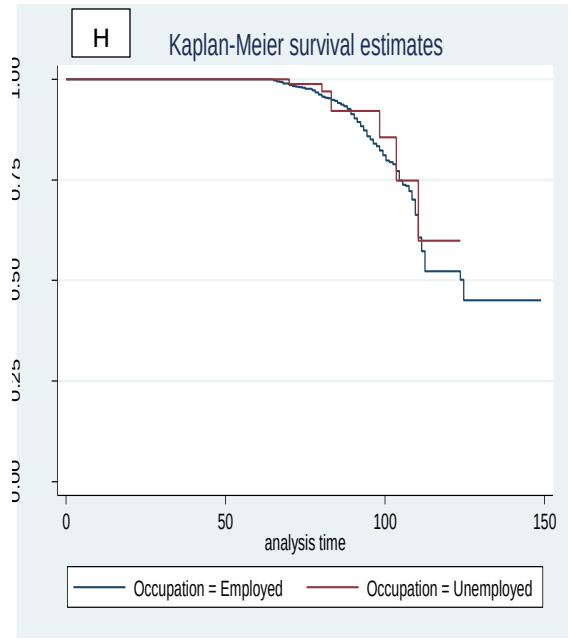
Figure7. 1 (A-J): Graphs of K-M Survivor Functions for all Categorical Covariates

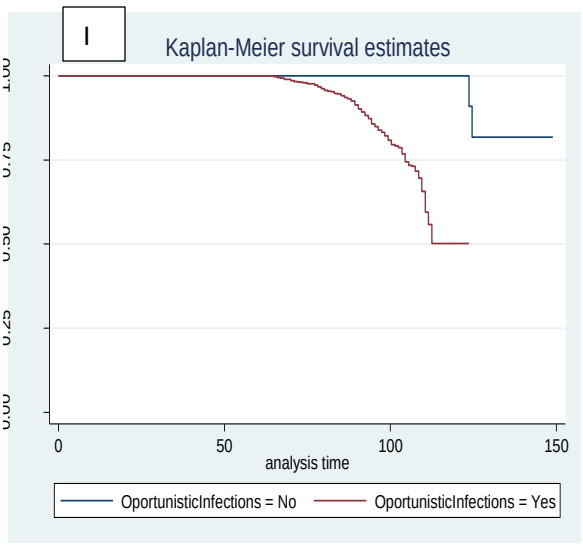



F
Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
50
100
150
analysis time
TBStatus = Negative
TBStatus = Positive


F
Kaplan-Meier survival estimates
0
50
100
150
analysis time
MaritalStatus = Never Married
MaritalStatus = Married
MaritalStatus = Widowed or Divorced


G
Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
50
100
150
analysis time
Education = Educated
Education = Not Educated


H
Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
50
100
150
analysis time
Occupation = Employed
Occupation = Unemployed


I
Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
50
100
150
analysis time
OportunisticInfections = No
OportunisticInfections = Yes

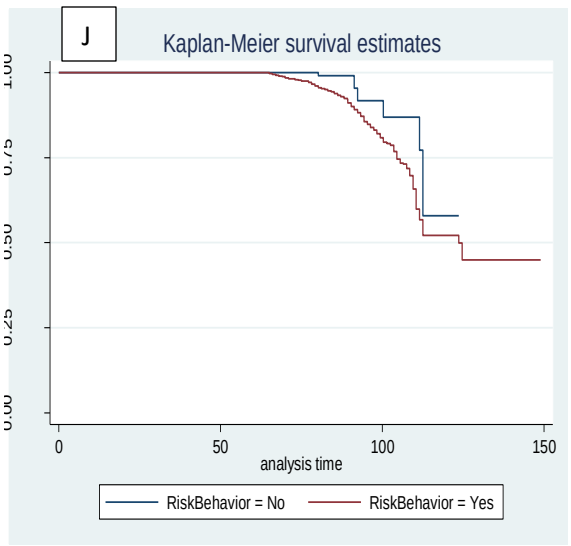

J
Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
50
100
150
analysis time
RiskBehavior = No
RiskBehavior = Yes

Figure7. 2 (A-G): Cumulative Hazard Plots for Testing PH Assumptions

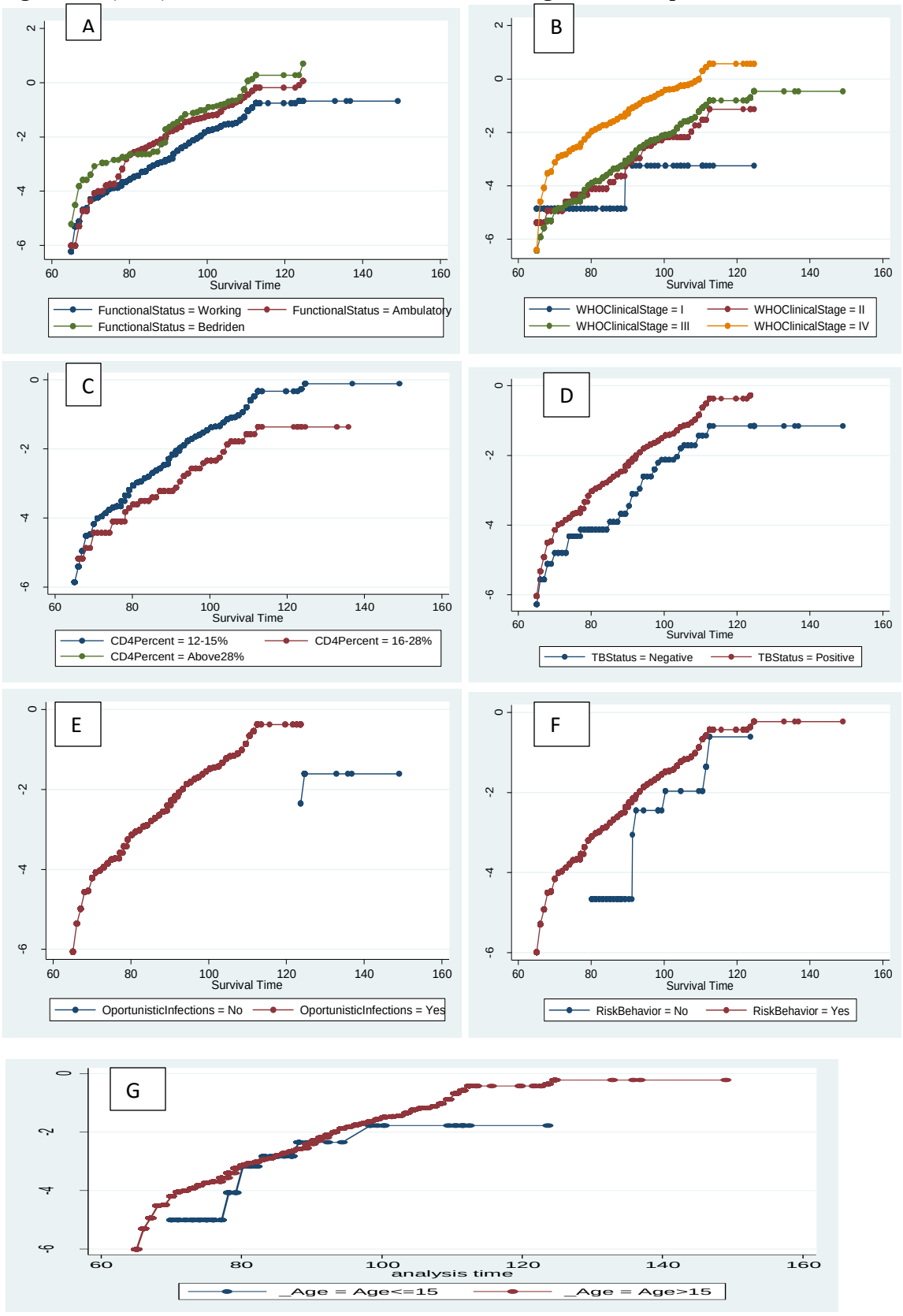


Table7. 2: Test of PH Assumption by Using Schoenfeld Residuals

Covariates	Rho	Chi2	DF	Prob> Chi2
WHO Clinical Stage (I as Reference)			1	
II	0.05811	1.74	1	0.1869
III	0.07151	2.64	1	0.1042
IV	0.05200	1.41	1	0.2357
CD4 Percent (12-15% as Reference)			1	
16-28%	-0.04816	1.22	1	0.2700
Above 28%			1	
TB Status (Negative as Reference)			1	
Positive	-0.01628	0.16	1	0.6899
Functional Status (Working as Reference)			1	
Ambulatory	0.02268	0.29	1	0.5871
Bedridden	0.06059	2.10	1	0.1473
Opportunistic Infections (No as Reference)			1	
Yes	-0.04282	0.85	1	0.3558
Risk Behavior (No as Reference)			1	
Yes	-0.06201	2.16	1	0.1416
Age	0.07981	4.05	1	0.0442

Figure7. 3: Test of PH Assumption by Using Plot of Scaled Schoenfeld Residuals for Age against Rank of Survival Time

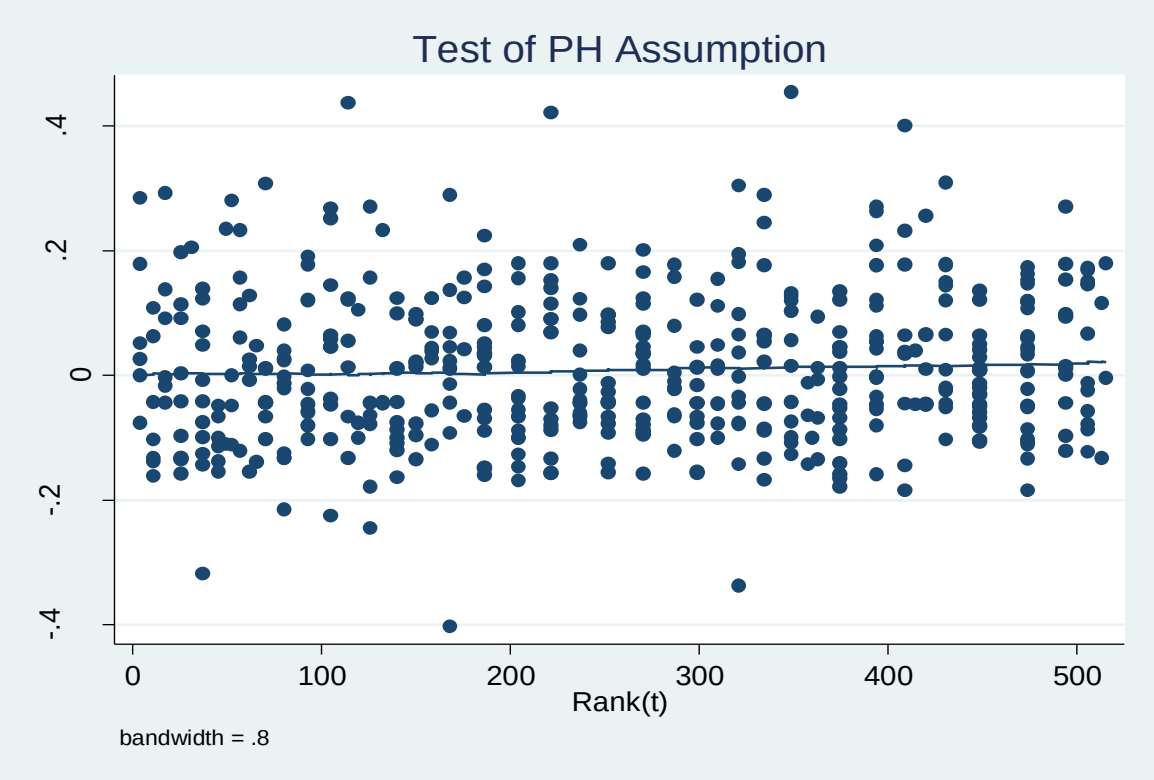
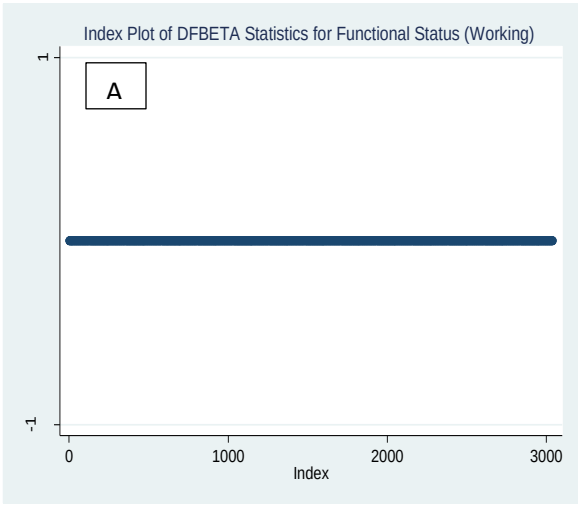
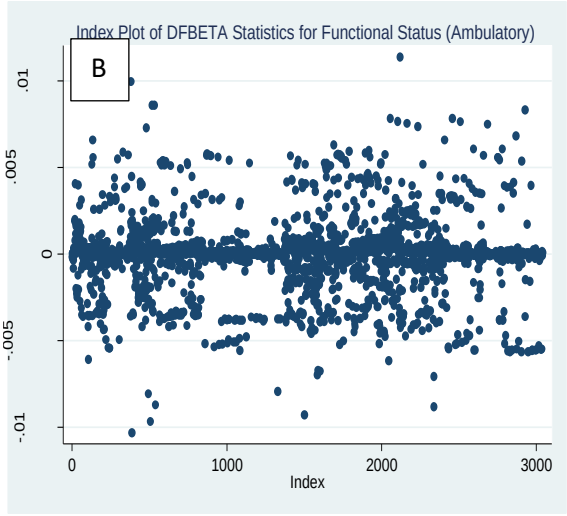
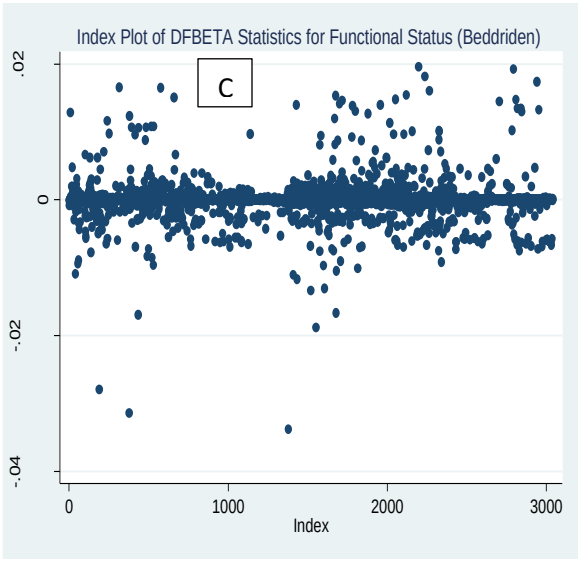
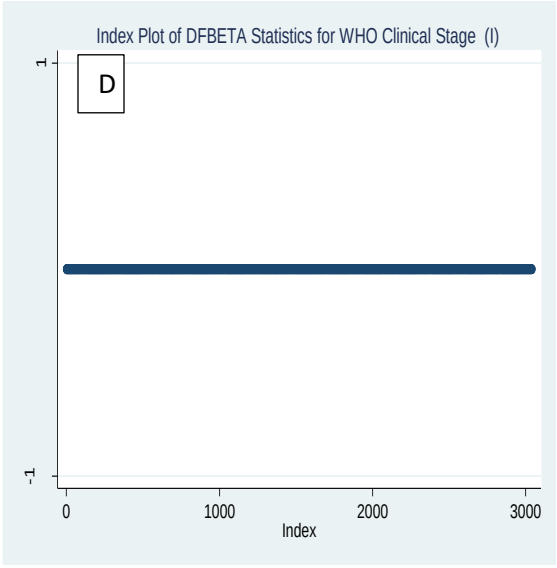


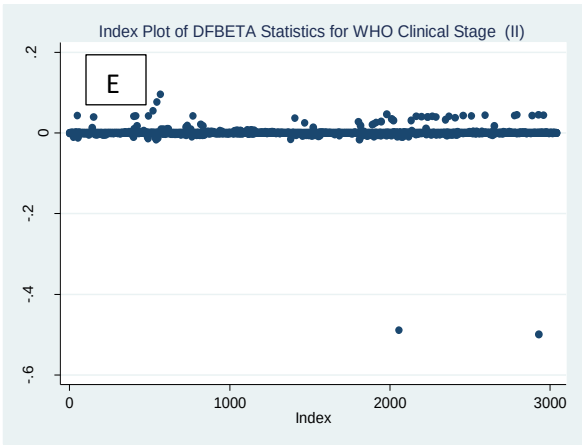
Figure7. 4 (A-Q): Index Plots of DFBETA Statistics for All Indicators in the Cox PH Model

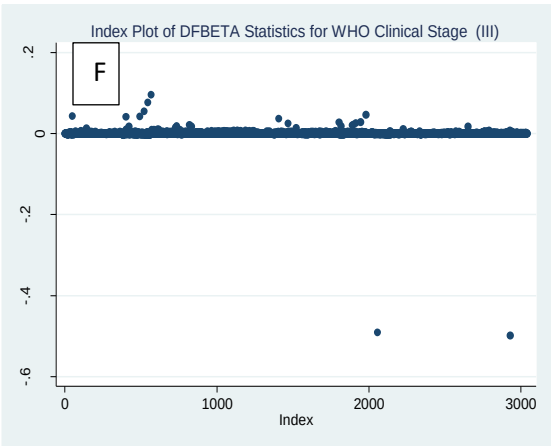


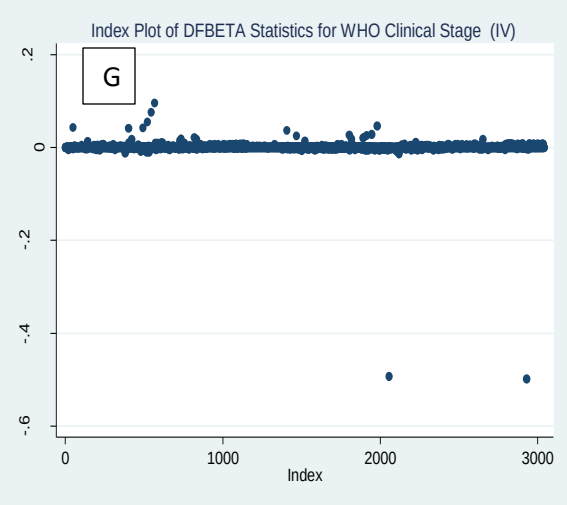


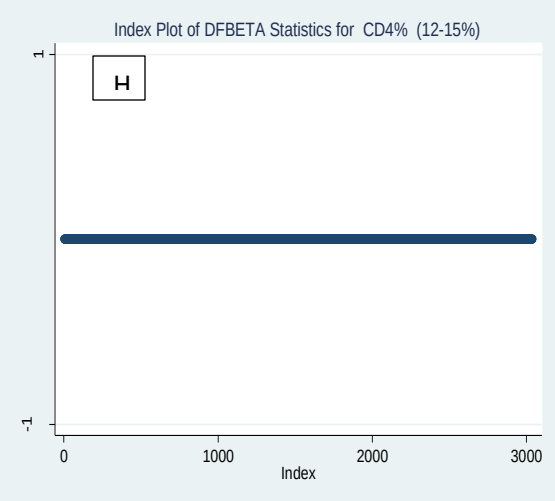


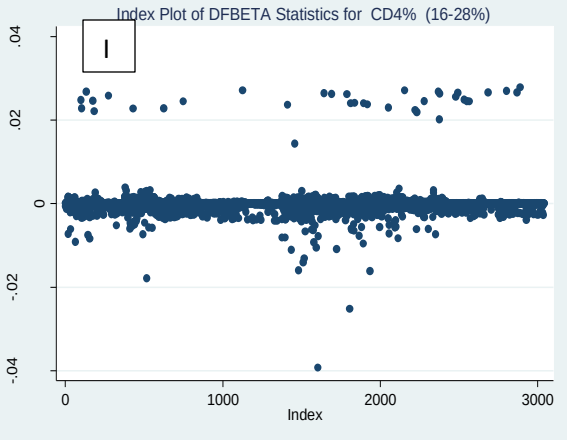


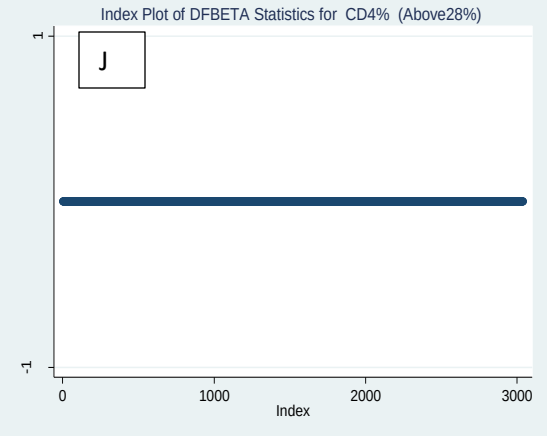


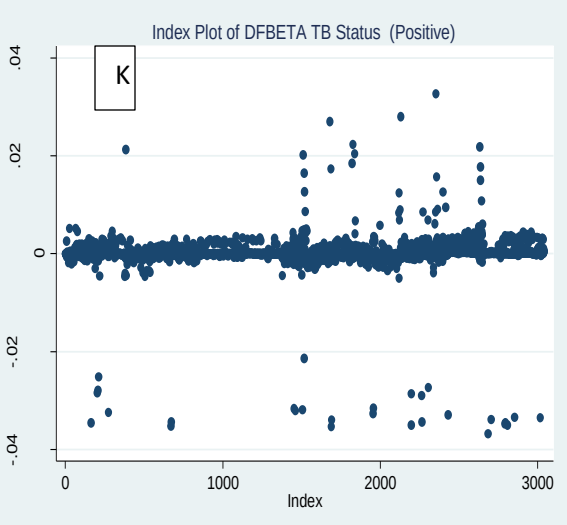


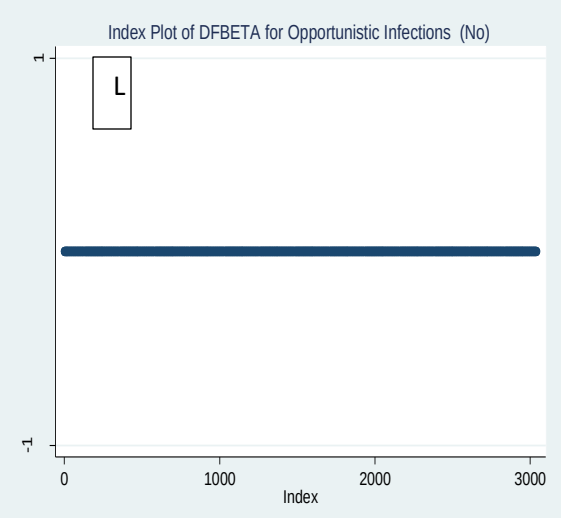

Index Plot of DFBETA Statistics for WHO Clinical Stage (IV)
G
.2
0
-.2
-.4
-.6
0
1000
2000
3000
Index

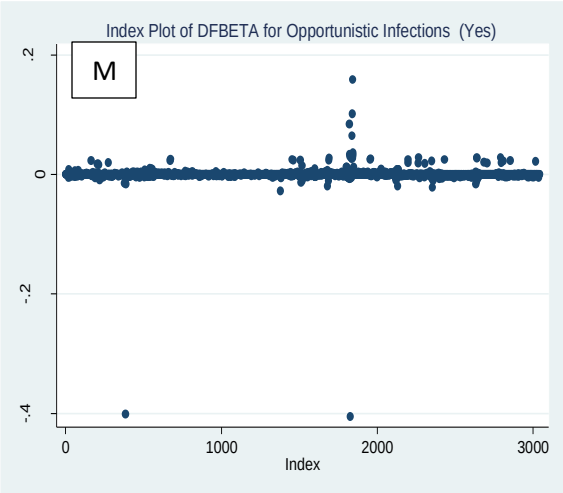

Index Plot of DFBETA Statistics for CD4% (12-15%)
H
1
-1
0
1000
2000
3000
Index

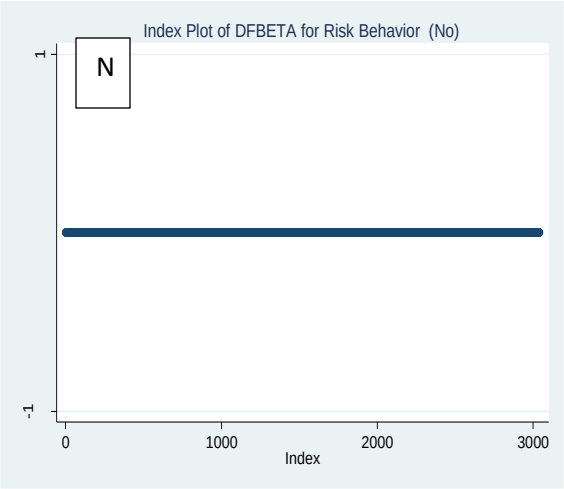

Index Plot of DFBETA Statistics for CD4% (16-28%)
I
.04
.02
0
-.02
-.04
0
1000
2000
3000
Index

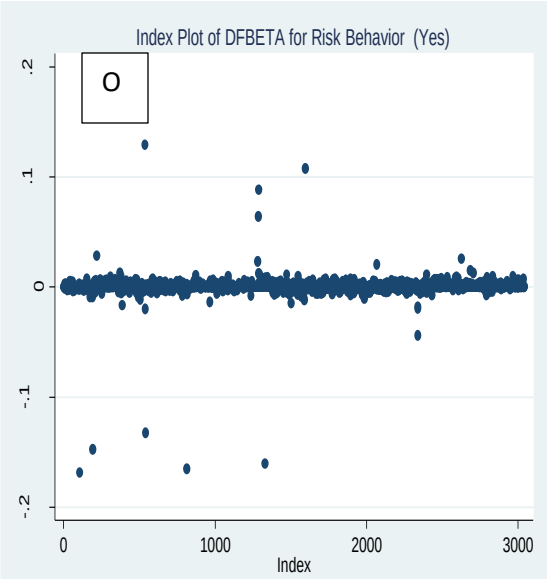

Index Plot of DFBETA Statistics for CD4% (Above28%)
J
1
-1
0
1000
2000
3000
Index

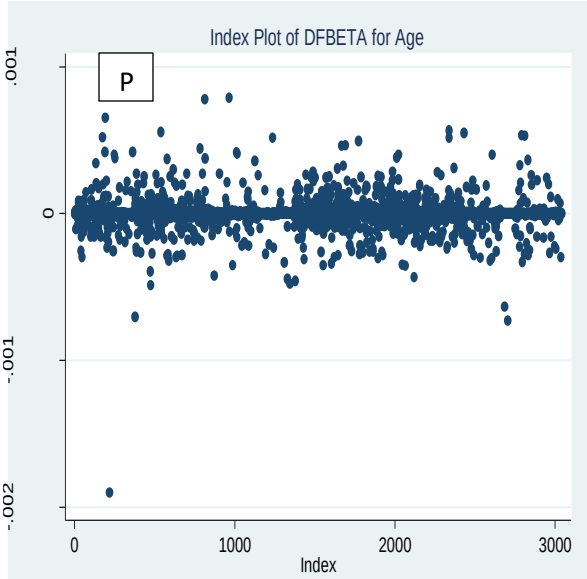

Index Plot of DFBETA TB Status (Positive)
K
.04
.02
0
-.02
-.04
0
1000
2000
3000
Index


Index Plot of DFBETA for Opportunistic Infections (No)
L
1
-1
0
1000
2000
3000
Index


Index Plot of DFBETA for Opportunistic Infections (Yes)
M
.2
0
-.2
-.4
0
1000
2000
3000
Index


Index Plot of DFBETA for Risk Behavior (No)
N
1
-1
0
1000
2000
3000
Index


Index Plot of DFBETA for Risk Behavior (Yes)
O
.2
.1
0
-.1
-.2
0
1000
2000
3000
Index


Index Plot of DFBETA for Age
P
.001
0
-.001
-.002
0
1000
2000
3000
Index

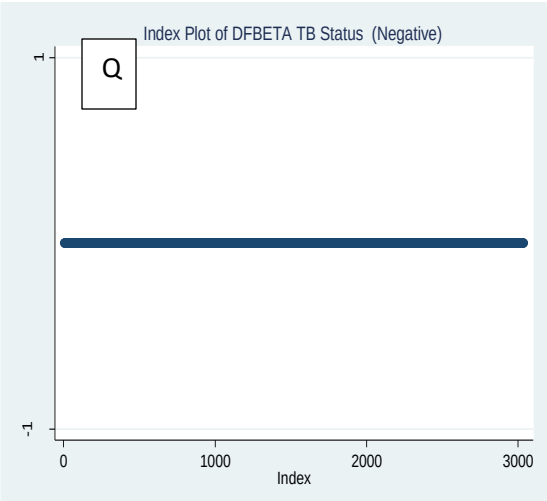

Index Plot of DFBETA TB Status (Negative)
Q
1
-1
0
1000
2000
3000
Index

Figure7. 5 (A-D): Quantile-Quantile Plots of Survival Time for Binary Covariates

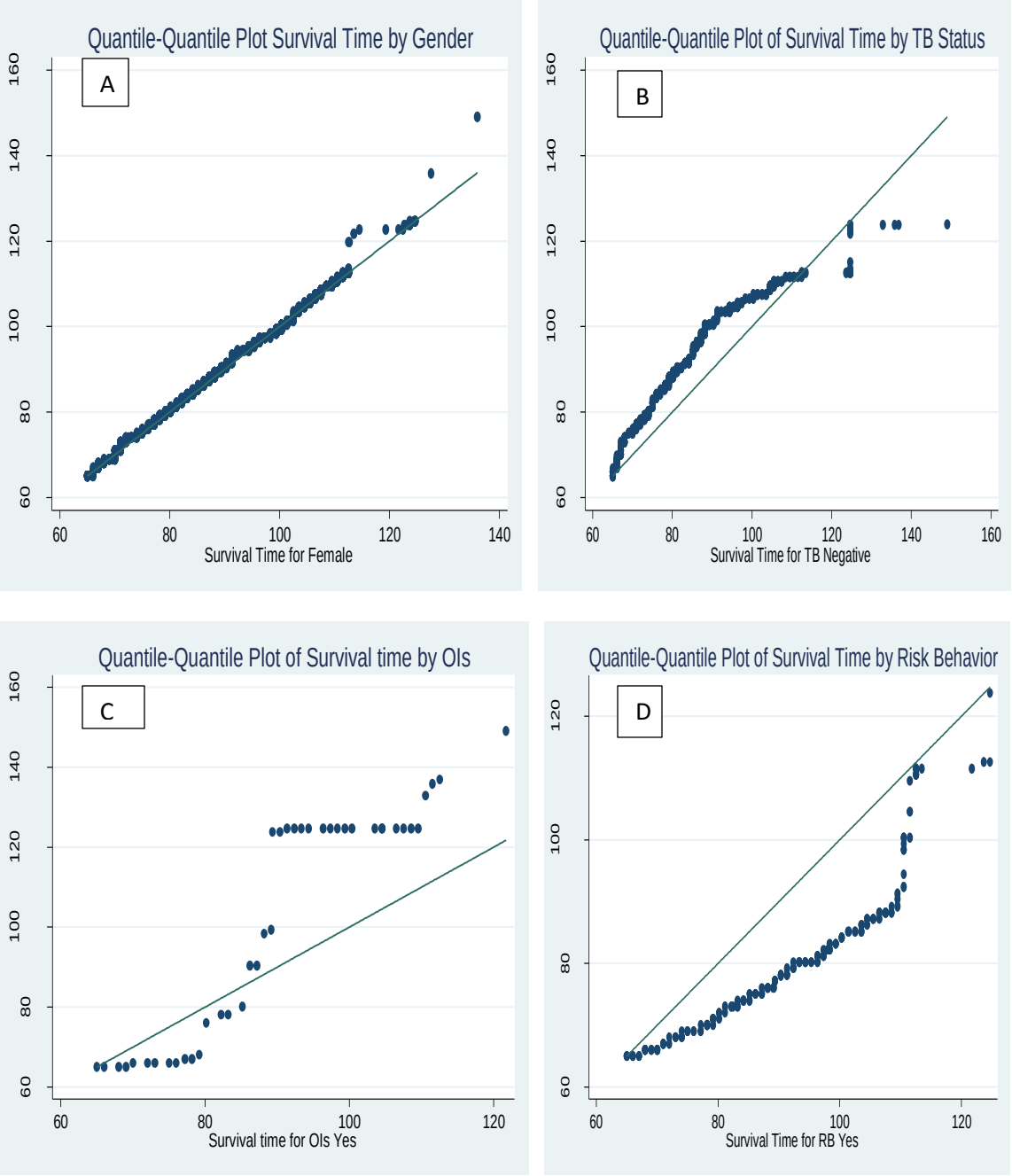


Table: 7. 3: Parametric PH Models Analysis

Covariates	Exponential PH				Weibull PH				Gompertz PH			
	β	p	95%CI FOR β		β	p	95%CI FOR β		β	p	95%CI FOR β	
Age	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.02
Gender	0.20	0.04	0.01	0.40								
FunctionalStatus												
1	0.32	0.00	0.12	0.52	0.25	0.01	0.05	0.45	0.26	0.01	0.07	0.46
2	0.32	0.03	0.04	0.61	0.15	0.31	-0.14	0.43	0.14	0.35	-0.15	0.42
Weight	-0.01	0.02	-0.03	-0.00	-0.01	0.04	-0.02	-0.00	-0.01	0.03	-0.02	-0.00
WHOClinicalStage												
1	1.12	0.13	-0.33	2.57	0.92	0.21	-0.52	2.37	0.93	0.21	-0.52	2.37
2	1.71	0.02	0.31	3.10	1.25	0.08	-0.15	2.64	1.24	0.08	-0.16	2.63
3	2.84	0.00	1.44	4.24	2.60	0.00	1.20	4.00	2.60	0.00	1.21	4.00
CD4Percent												
2	-0.75	0.00	-1.09	-0.41	-0.67	0.00	-1.01	-0.33	-0.67	0.00	-1.01	-0.33
3	-14.6	0.99	-1577.1	1548.0	-14.03	0.98	-1049.6	1021.6	-13.52	0.97	-810.9	783.85
1.TBStatus	1.11	0.00	0.74	1.48	0.73	0.00	0.34	1.11	0.72	0.00	0.34	1.11
Oportunistic Infect~s					1.38	0.01	0.35	2.41	1.68	0.00	0.65	2.72
1.RiskBehavior	1.57	0.00	0.70	2.44	0.99	0.02	0.16	1.82	1.03	0.02	0.19	1.86
_cons	-10.69	0.00	-12.34	-9.04	-43.13	0.00	-46.34	-39.93	-17.3	0.00	-19.4	-15.31
/ln_p					2.08	0.00	2.02	2.15				
P					8.03		7.51	8.59				
1/p					0.12		0.12	0.13				
/gamma									0.08	0.00	0.08	0.09